STATISTICAL METHODS AND THEORY FOR ANALYZING HIGH DIMENSIONAL TIME SERIES

by

Huitong Qiu

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2016

Abstract

High dimensional time series[1] presents uniques challenges due to both the serial dependence and the large feature space. In this research, we consider three topics under high dimensional time series: graphical model estimation under multiple time series, portfolio optimization under heavy-tailed time series, and Kolmogorov dependent time series. In the first topic, we consider multiple stationary time series with varying covariance structure, and propose a graphical model estimator that borrows strength from all time series. In the second topic, we consider financial asset return series that exhibit heavy-tailed distributions. We reformulate portfolio optimization based on quantile statistics to explicitly accommodate heavy tails. In the third topic, we propose a general framework for modeling serial dependence in multivariate time series. We explore its connections with existing models, and demonstrate its applications in scatter matrix estimation. At the core of these topics are several methods for estimating high dimensional covariance and scatter matrices, and the quantification of how their consistency is affected by the dependence strength of the time series.

[1] Detailed definition is introduced in Chapter 1.

ABSTRACT

Advisor:

Brian Caffo, PhD

Committee:

Michelle Carlson, PhD (Committee Chair, SPH Mental Health & Epidemiology)

Brian Caffo, PhD (Thesis Advisor, SPH Biostatistics)

Martin Lindquist, PhD (SPH Biostatistics)

Tamas Budavari, PhD (EN Applied Math and Statistics)

Fang Han, PhD (SPH Biostatistics)

Alternates:

Vadim Zipunnikov, PhD (SPH Biostatistics)

Jonathan Links, PhD (SPH Environmental Health Sciences & Health Policy and Management)

Acknowledgments

First and foremost I would like to express my deepest thanks to my advisor and mentor, Brian Caffo, for his guidance, encouragement, and friendship. He has been a constant source of ideas that have led me to the greatest excitements of statistical research. He has been a first resort for advice and support when I get lost in my life and career. And his nice and amiable personality has made my research a joyful experience. I would also like to give my special thanks to Fang Han, who has been a close collaborator and a great friend. His vision in statistics and machine learning has provided invaluable guidance to my PhD study. And his insights have inspired a number of works in my research.

I am very grateful for my thesis and oral exam committee for their advice and criticism. They are Profs. Michelle Carlson, Brian Caffo, Martin Lindquist, Tamas Budavari, Fang Han, Derek Cummings, Daniel Robinson, Vadim Zipunnikov, and Jonathan Links.

I thank the department of biostatistics for providing an open, collaborative, and inspiring environment for research. Special thanks to all the faculty members, including Ciprian Crainiceanu, Michael Rosenblum, Daniel Sharfstein, Jeff Leek, Constantine Fragakis, and Mei-cheng Wang, for their great teaching and advising on various projects. Thanks also to

ACKNOWLEDGMENTS

Han Liu from Princeton University for his guidance on a number of my works.

Thanks to the students in our department for making it a fun community. Special thanks to Shaojie Chen, Chen Yue, Lei Huang, Detian Deng, Yuting Xu, and Sheng Xu. I enjoyed our adventures over the past five years. Thanks also to my seniors, including Haochang Shou, Zhenke Wu, Yingying Wei, Juemin Yang, Yifei Sun, Yi Lu, Shanshan Li, and Yang Ning, for their friendship, help, and advice.

Finally, I would like to express my sincere thanks to my family. Thanks to my parents for their selfless love and support. Thanks to my brother, Huida Qiu, for sharing all my happiness and sorrows, and for being my ultimate resort for help and strength. It has been a privilege to be part of this great family. Special thanks to Yuan Chen, my wife, for her support, care, encouragement, and all the happiness we have enjoyed together.

Contents

Abstract **ii**

Acknowledgments **iv**

List of Tables **xi**

List of Figures **xii**

1 Introduction **1**

1.1 Organization 3

2 Joint Estimation of Graphical Models under Multiple Time Series **6**

2.1 Introduction 7

2.2 The Model and Method 11

2.2.1 Model 12

2.2.2 Method 14

2.3 Theoretical Properties 17

2.4 Experiments 23
2.4.1 Synthetic Data 23
2.4.1.1 Setting 1: Simultaneously Evolving Edges 26
2.4.1.2 Setting 2: Sequentially Growing Edges 28
2.4.1.3 Setting 3: Random Edges 29
2.4.2 Impact of a Small Label Size n 30
2.4.3 ADHD-200 Data 31
2.4.3.1 Brain Development 32
2.5 Discussion 35

3 Robust Portfolio Optimization **36**

3.1 Introduction 37
3.2 Background 39
3.2.1 Notation 39
3.2.2 Gross-exposure Constrained Global Minimum Variance Formulation 40
3.3 Method 41
3.4 Theoretical Properties 44
3.5 Experiments 51
3.5.1 Synthetic Data 52
3.5.2 Real Data 55
3.6 Discussion 56

4 A Theory of Kolmogorov Dependence with Applications to Scatter Matrix Estimation 58

4.1 Introduction 59

4.1.1 Organization 63

4.1.2 Notation 64

4.2 Kolmogorov Dependence 65

4.3 Robust Scatter Matrix Estimation 74

4.4 Proof of Main Results 79

4.4.1 Proof of Results in Section 4.2 79

4.4.2 Proof of Results in Section 4.3 87

4.5 Discussion 94

5 Discussion and Future Work 96

Appendices 99

A Appendix to Chapter 2 99

A.1 Auto-Correlation and Cross-Correlation 99

A.2 Additional Experiments 102

A.2.1 Impact of Temporal Dependence 102

A.2.2 Impact of Label Size n, Sample Size T, and Dimension d. 104

A.2.3 Additional Results on ADHD-200 Data 105

A.2.3.1 Development of Brain Network Density 105

A.2.3.2 The Impact of Bandwidth . . . 106
A.3 Techinical Proofs . . . 107
A.3.1 Proof of Lemma 1 . . . 107
A.3.1.1 Proof of Lemma 1 . . . 114
A.3.2 Proof of Theorem A.1.1 . . . 116
A.3.2.1 Proof of Theorem A.1.1 . . . 117
A.3.3 Proof of Lemma 2 . . . 117
A.3.3.1 Proof of Lemma 2 . . . 119

B Appendix to Chapter 3 **121**

B.1 Supporting Lemmas . . . 121
B.2 Proofs of the Main Results . . . 132
B.2.1 Proof of Lemma 3 . . . 132
B.2.2 Proof of Theorem 6 . . . 132
B.2.3 Proof of Theorem 9 . . . 136
B.3 Matrix Projection . . . 138

C Appendix to Chapter 4 **142**

C.1 Concentration Inequalities under Weak Dependence . . . 142
C.2 Supporting Lemma . . . 148

Bibliography **155**

CONTENTS

Curriculum Vitae **172**

List of Tables

2.1 Comparison of inverse covariance matrix estimation errors in three data generating models. The parameter estimation error with regard to the matrix ℓ_1, ℓ_2, and Frobenius norms (denoted as ℓ_F here) is provided with standard deviations in parentheses. The results are obtained by 1,000 simulations. 30

3.1 Parameters for generating the covariance matrix in Equation (3.15). 53
3.2 Annualized Sharpe ratios, returns, and risks under 4 competing approaches, using S&P 500 index data. 55

4.1 Important examples of weak dependence. 72

List of Figures

2.1 ROC curves of four competing methods under three settings: simultaneous (a-e), sequential (f-i), and random (j). The target labels are $u_0 = 0$ except for in (c), where $u_0 = 1/2$. In each setting we set the dimension $d = 50$, the number of labels $n = 51$, the number of observations $T = 100$, and the result is obtained by 1,000 simulations. 27
2.2 ROC curves of KSE and naive under Setting 1: sequentially evolving edges. We set dimension $d = 50$; number of labels $n = 3$; number of pre-fixed edges $n_{\text{fix}=100}$; number of growing edges $n_{\text{grow}} = 500$. 32
2.3 Estimated brain connectivity network at ages 7.09, 11.75, 21.83 in healthy subjects. 34

3.1 Portfolio risks, selected number of stocks, and matching rates to the oracle optimal portfolios. 53

Chapter 1

Introduction

A multivariate time series is a sequence of random vectors observed successively over a time interval. As a characteristic property, the random vectors in the time series often exhibit serial dependence. In particular, the value of a random vector at one time is statistically dependent on the value at another time. In this work, we consider high dimensional time series where the dimension of the random vectors can be much larger than the number of observations.

High dimensional time series arise in a wide spectrum of scientific applications. For example, in brain functional magnetic resonance imaging (fMRI), the image from one scan is highly dependent on the images from neighboring scans. Moreover, there are usually hundreds of thousands of voxels in an image, while the number of repeated scans is often only a few hundred for a subject. In finance, the current prices of the stocks in a portfolio are highly dependent on the historical price movements. Moreover, since the market conditions change rapidly, the number of price observations that reflect the current market conditions are often much smaller than the number of stocks in a portfolio.

High dimensional time series present unique challenges in statistical analysis. First of all, quantifying the degree of serial dependence is difficult. Although many quantifications exists, they are mostly tailored to specific models and methods, and are not immediately applicable to others. Moreover, the connections between these quantifications are largely unknown. Secondly, serial dependence violates the assumption of independent observations in classic statistical analysis. How to characterize the effect of serial dependence in statistical estimation is still an open question in many applications. Thirdly, to accommo-

date a much larger dimension compared to sample size, special regularization techniques are needed to reduce the feature space.

In this work, we tackle these challenges in three specific topics of high dimensional time series: estimating graphical models in multiple time series, optimizing portfolios under heavy-tailed financial asset return series, and modeling the serial dependence strength of a general time series. Detailed specifications and contributions under each topic follow in Section 1.1. At the core of the proposed methodologies are high dimensional covariance or scatter matrix estimators. A common theme of the proposed theory is to quantify of how serial dependence impacts the consistency of these methods.

1.1 Organization

In the first part of the thesis, we consider the problem of jointly estimating multiple graphical models[1] in multiple time series. Motivated by a resting state functional magnetic resonance imaging (rs-fMRI) study, we consider data collected from n subjects, each of which consists of T stationary but dependent observations. The distributions of the data vary across subjects, but are assumed to change smoothly corresponding to a measure of closeness between subjects. In this scenario, statistical methodologies are desired to estimate the graphical model of any distribution, while borrowing strength from all the subjects available. To this end, we propose a kernel based method for estimating the covariance ma-

[1] A graphical model is a statistical model whose conditional dependence structure is represented by a graph. The nodes of the graph represent random variables, and the edges represent the conditional dependence structure between the random variables.

trices and graphical models. Theoretically, under a double asymptotic framework, where both (T, n) and the dimension d can increase, we provide the explicit rate of convergence in parameter estimation. It characterizes the strength one can borrow across different individuals and the impact of serial dependence on parameter estimation. Empirically, experiments on both synthetic and real rs-fMRI data illustrate the effectiveness of the proposed method.

The second part of the thesis is focused on portfolio optimization under financial asset return series. Financial asset returns typically exhibit heavy-tailed distributions[2], where significant deviations from the mean is far more likely to occur than in Gaussian distributions. Heavy-tailed distributions make the modeling and analysis of financial returns challenging, since many standard, moment-based statistics are no longer consistent, or even ill-defined, without light-tail assumptions. In this work, we consider a stationary, high dimensional time series with no assumption on the tail condition. We propose a robust portfolio optimization approach building on a class of quantile-based scatter matrix estimators. We derive explicit rates of convergence for the scatter matrix estimators and the risk of the optimized portfolio. The rates capture the effect of serial dependence, measured by ϕ-mixing coefficients, on consistency, and hold without any requirement on the tail of the distributions. The empirical effectiveness of the proposed method is demonstrated under both synthetic and real equity data.

In the third part of the thesis, we develop a general framework for modeling serial dependence for time series. The framework is motivated by the difficulty of using existing

[2]Heavy-tailed distributions commonly refer to probability distributions whose tails cannot be upper bounded by the exponential distribution.

models to analyze quantile-based statistics, as well as the lack of unity over existing models. To these ends, we propose a new measure of serial dependence named Kolmogorov dependence measure. Using this measure, we develop the Kolmogorov dependence condition, and show that it's weaker and more intuitive than many widely used weak dependence conditions. Under the framework of Kolmogorov dependence, we revisit the topic of estimating quantile-based scatter matrices. We show that a more general characterization of the effect of dependence on the consistency can be obtained.

Chapter 2

Joint Estimation of Graphical Models under Multiple Time Series

2.1 Introduction

Undirected graphical models encoding the conditional independence structure among the variables in a random vector have been heavily exploited in multivariate data analysis (Lauritzen, 1996). For a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathsf{T}}$, the corresponding undirected graphical model specifies a graph with node set $\mathrm{V} = \{1, \ldots, d\}$ and edge set $\mathrm{E} = \{(i, j) : X_i \text{ and } X_j \text{ are conditional dependent given the remaining random variables in } \boldsymbol{X}\}$. In particular, when $\boldsymbol{X} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ is multivariate Gaussian, estimating such graphical models is equivalent to estimating the nonzero entries in the inverse covariance matrix $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$. Indeed, the edge set is equal to $\mathrm{E} = \{(i, j) : \boldsymbol{\Theta}_{ij} \neq 0\}$ (Dempster, 1972). The undirected graphical model encoding the conditional independence structure for the Gaussian distribution is sometimes called a Gaussian graphical model.

There has been much work on estimating a single Gaussian graphical model, $\mathbf{G}$, based on n independent observations. In low dimensional settings where the dimension, d, is fixed, Drton and Perlman (2007) and Drton and Perlman (2008) proposed to estimate $\mathbf{G}$ using multiple testing procedures. In settings where the dimension is much larger than the sample size, n, Meinshausen and Bühlmann (2006) proposed to estimate $\mathbf{G}$ by solving a collection of regression problems via the lasso. Yuan and Lin (2007), Banerjee et al. (2008), Friedman et al. (2008), Rothman et al. (2008), and Liu and Luo (2012) proposed to directly estimate $\boldsymbol{\Theta}$ using the ℓ_1 penalty (detailed definition provided later). More recently, Yuan (2010) and Cai et al. (2011) proposed to estimate $\boldsymbol{\Theta}$ via linear programming. The above mentioned estimators are all consistent with regard to both parameter estimation and

model selection, even when d is nearly exponentially larger than n.

This body of work is focused on estimating a single graph based on independent realizations of a common random vector. However, in many applications this simple model does not hold. For example, the data can be collected from multiple individuals that share the same set of variables, but differ with regard to the structures among variables. This situation is frequently encountered in the area of brain connectivity network estimation (Friston, 2011). Here brain connectivity networks corresponding to different subjects vary, but are expected to be more similar if the corresponding subjects share many common demographic, health or other covariate features. Under this setting, estimating the graphical models separately for each subject ignores the similarity between the adjacent graphical models. In contrast, estimating one population graphical model based on the data of all subjects ignores the differences between graphs and may lead to inconsistent estimates.

There has been a line of research in jointly estimating multiple Gaussian graphical models for independent data. On one hand, Guo et al. (2011) and Danaher et al. (2014) proposed methods via introducing new penalty terms, which encourage the sparsity of both the parameters in each subject and the differences between parameters in different subjects. On the other hand, Song et al. (2009a), Song et al. (2009b), Kolar and Xing (2009), Zhou et al. (2010), and Kolar et al. (2010) focused on independent data with time-varying networks. They proposed efficient algorithms for estimating and predicting the networks along the time line.

In this paper, we propose a new method for jointly estimating and predicting networks

corresponding to multiple subjects. The method is based on a different model compared to the ones listed above. The motivation of this model arises from resting state functional magnetic resonance imaging (rs-fMRI) data, where there exist many natural orderings corresponding to measures of health status, demographics, and many other subject-specific covariates. Moreover, the observations of each subject are multiple brain scans with temporal dependence. Accordingly, different from the methods in estimating time varying networks, we need to handle the data where each subject has T, instead of one, observations. Different from the methods in Guo et al. (2011) and Danaher et al. (2014), it is assumed that there exists a natural ordering for the subjects, and the parameters of interest vary smoothly corresponding to this ordering. Moreover, we allow the observations to be dependent via a temporal dependence structure. Such a setting has not been studied in high dimensions until very recently (Loh and Wainwright, 2012; Han and Liu, 2013b; Wang et al., 2013).

We exploit a similar kernel based approach as in Zhou et al. (2010). It is shown that our method can efficiently estimate and predict multiple networks while allowing the data to be dependent. Theoretically, under a double asymptotic framework, where both d and (T, n) may increase, we provide an explicit rate of convergence in parameter estimation. It sharply characterizes the strength one can borrow across different subjects and the impact of data dependence on the convergence rate. Empirically, we illustrate the effectiveness of the proposed method on both synthetic and real rs-fMRI data. In detail, we conduct comparisons of the proposed approach with several existing methods under three synthetic patterns of evolving graphs. In addition, we study the large scale ADHD-200 dataset to

investigate the development of brain connectivity networks over age, as well as the effect of kernel bandwidth on estimation, where scientifically interesting results are unveiled.

We note that the proposed multiple time series model has analogous prototypes in spatial-temporal analysis. This line of work is focused on multiple times series indexed by a spatial variable. A common strategy models the spatial-temporal observations by a joint Gaussian process, and imposes a specific structure on the spatial-temporal covariance function (Jones and Zhang, 1997; Cressie and Huang, 1999). Another common strategy decomposes the temporal series into a latent spatial-temporal structure and a residual noise. Examples of the latent spatial-temporal structure include temporal autoregressive processes (Høst et al., 1995; Sølna and Switzer, 1996; Antunes and Rao, 2006; Rao, 2008) and mean processes (Storvik et al., 2002; Gelfand et al., 2003; Banerjee et al., 2004, 2008; Nobre et al., 2011). The residual noise is commonly modeled by a parametric process such as a Gaussian process. The aforementioned literature is restricted in three aspects. First, they only consider univariate or low dimensional multivariate spatial-temporal series. Secondly, they restrict the covariance structure of the observations to a specific form. Thirdly, none of this literature addresses the problem of estimating the conditional independence structure of the time series. In comparison, we consider estimating the conditional independence graph under high dimensional times series. Moreover, our model involves no assumption on the structure of the covariance matrix.

We organize the rest of the paper as follows. In Section 2.2, the problem setup is introduced and the proposed method is given. In Section 2.3, the main theoretical results

are provided. In Section 2.4, the method is applied to both synthetic and rs-fMRI data to illustrate its empirical usefulness. A discussion is provided in the last section. Additional results and technical proofs are put in the appendix.

2.2 The Model and Method

Let $\mathbf{M} = (M_{jk}) \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$. We denote $\boldsymbol{v}_I$ to be the sub-vector of $\boldsymbol{v}$ whose entries are indexed by a set $I \subset \{1, \ldots, d\}$. We denote $\mathbf{M}_{I,J}$ to be the submatrix of $\mathbf{M}$ whose rows are indexed by I and columns are indexed by J. Let $\mathbf{M}_{I,*}$ be the submatrix of $\mathbf{M}$ whose rows are indexed by I, and $\mathbf{M}_{*,J}$ be the submatrix of $\mathbf{M}$ whose columns are indexed by J. For $0 < q < \infty$, define the ℓ_0, ℓ_q, and ℓ_∞ vector norms as

$$\|\boldsymbol{v}\|_0 = \sum_{j=1}^{d} I(v_j \neq 0), \;\; \|\boldsymbol{v}\|_q := \Big(\sum_{j=1}^{d} |v_j|^q\Big)^{1/q}, \;\; \text{and} \;\; \|\boldsymbol{v}\|_\infty = \max_{1\leq j\leq d} |v_j|,$$

where $I(\cdot)$ is the indicator function. For a matrix $\mathbf{M}$, denote the matrix ℓ_q, $\ell_{\max}$, and Frobenius norms to be

$$\|\mathbf{M}\|_q = \max_{\|\boldsymbol{v}\|_q=1} \|\mathbf{M}\boldsymbol{v}\|_q, \;\; \|\mathbf{M}\|_{\max} = \max_{jk} |M_{jk}|, \;\; \text{and} \;\; \|\mathbf{M}\|_F = \left(\sum_{j,k} |M_{jk}|^2\right)^{1/2}.$$

For any two sequences $a_n, b_n \in \mathbb{R}$, we say that $a_n \asymp b_n$ if $cb_n \leq a_n \leq Cb_n$ for some constants c, C.

2.2.1 Model

Let $\{\boldsymbol{X}^u\}_{u\in[0,1]}$ be a series of d-dimensional random vectors indexed by the label u, which can represent any kind of ordering in subjects (e.g., any covariate or confounder of interest transformed to the space $[0,1]$). For any $u \in [0,1]$, assume that $\boldsymbol{X}^u \sim N_d\{\mathbf{0}, \boldsymbol{\Sigma}(u)\}$. Here $\boldsymbol{\Sigma}(\cdot) : [0,1] \to \mathbb{S}_+^{d\times d}$ is a function from $[0,1]$ to the d by d positive definite matrix set, $\mathbb{S}_+^{d\times d}$. Let $\boldsymbol{\Omega}(u) := \{\boldsymbol{\Sigma}(u)\}^{-1}$ be the inverse covariance matrix of $\boldsymbol{X}^u$ and let $\mathbf{G}(u) \in \{0,1\}^{d\times d}$ represent the conditional independence graph corresponding to $\boldsymbol{X}^u$, satisfying that $\{\mathbf{G}(u)\}_{jk} = 1$ if and only if $\{\boldsymbol{\Omega}(u)\}_{jk} \neq 0$.

Suppose that data points in $u = u_1, \ldots, u_n$ are observed. Let $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT} \in \mathbb{R}^d$ be T observations of $\boldsymbol{X}^{u_i}$, with a temporal dependence structure among them. In particular, for simplicity, in this manuscript we assume that $\{\boldsymbol{x}_{it}\}_{t=1}^T$ follows a lag one stationary vector autoregressive (VAR) model, i.e.,

$$\boldsymbol{x}_{it} = \mathbf{A}(u_i)\boldsymbol{x}_{i(t-1)} + \boldsymbol{\epsilon}_{it}, \text{ for } i = 1, \ldots, n, \ t = 2, \ldots, T, \tag{2.1}$$

and $\boldsymbol{x}_{it} \sim N_d\{\mathbf{0}, \boldsymbol{\Sigma}(u_i)\}$ for $t = 2, \ldots, T$. Here we note that extensions to vector autoregressive models with higher orders are also analyzable using the same techniques in Han and Liu (2013b). But for simplicity, in this manuscript we only consider the lag one case. $\mathbf{A}(u) \in \mathbb{R}^{d\times d}$ is referred to as the transition matrix. It is assumed that the Gaussian noise, $\boldsymbol{\epsilon}_{it} \sim N_d\{\mathbf{0}, \boldsymbol{\Psi}(u_i)\}$ is independent of $\{\boldsymbol{\epsilon}_{it'}\}_{t'\neq t}$ and $\{\boldsymbol{x}_{it'}\}_{t'=1}^{t-1}$. Both $\mathbf{A}(\cdot)$ and $\boldsymbol{\Psi}(\cdot)$ are considered as functions on $[0,1]$. Due to the stationary property, for any $u \in [0,1]$, taking the

covariance on either side of Equation (2.1), we have $\mathbf{\Sigma}(u) = \mathbf{A}(u)\mathbf{\Sigma}(u)\{\mathbf{A}(u)\}^{\mathsf{T}} + \mathbf{\Psi}(u)$. For any $i \neq i'$, it is assumed that $\{\boldsymbol{x}_{it}\}_{t=1}^{T}$ are independent of $\{\boldsymbol{x}_{i't}\}_{t=1}^{T}$. For $i = 1, \ldots, n$ and $t = 1, \ldots, T$, denote $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{itd})^{\mathsf{T}}$.

Of note, the function $\mathbf{A}(\cdot)$ characterizes the temporal dependence in the time series. For each label u, $\mathbf{A}(u)$ represents the transition matrix of the VAR model specific to u. By allowing $\mathbf{A}(u)$ to depend on u, as u varies, the temporal dependence structure of the corresponding time series is allowed to vary, too.

As is noted in Section 1, the proposed model is motivated by brain network estimation using rs-fMRI data. For instance, the ADHD data considered in Section 2.4.3 consist of n subjects with ages (u) ranging from 7 to 22, while time series measurements within each subject are indexed by t varying from 1 to 200, say. That is, for each subject, a list of rs-fMRI images with temporal dependence are available. We model the list of images by a VAR process, as exploited in Equation (2.1). For a fixed age u, $\mathbf{A}(u)$ characterizes the temporal dependence structure of the time series corresponding to the subject with age u. As age varies, the temporal dependence structures of the images may vary, too. Allowing $\mathbf{A}(u)$ to change with u accommodates such changes. The VAR model is a common tool in modeling dependence for rs-fMRI data. Consider Harrison et al. (2003), Penny et al. (2005), Rogers et al. (2010), Chen et al. (2011a), and Valdés-Sosa et al. (2005), for more details.

2.2.2 Method

We exploit the idea proposed in Zhou et al. (2010) and use a kernel based estimator for subject specific graph estimation. The proposed approach requires two main steps.. In the first step, a smoothed estimate of the covariance matrix $\mathbf{\Sigma}(u_0)$, denoted as $\mathbf{S}(u_0)$, is obtained for a target label u_0. In the second step, $\mathbf{\Omega}(u_0)$ is estimated by plugging the covariance matrix estimate $\mathbf{S}(u_0)$ into the CLIME algorithm (Cai et al., 2011).

More specifically, let $K(\cdot) : \mathbb{R} \to \mathbb{R}$ be a symmetric nonnegative kernel function with support set $[-1, 1]$. Moreover, for some absolute constant C_1, let $K(\cdot)$ satisfy that:

$$\sup_v K(v) \leq C_1, \quad \int_{-1}^{1} K(v)dv = 1, \quad \text{and} \quad \int_0^1 vK(v)dv \leq C_1. \tag{2.2}$$

Equation (2.2) is satisfied by a number of commonly used kernel functions. Examples include:

Uniform kernel: $K(s) = I(|s| \leq 1)/2$;

Triangular kernel: $K(s) = (1 - |s|)I(|s| \leq 1)$;

Epanechnikov kernel: $K(s) = 3(1 - s^2)I(|s| \leq 1)/4$;

Cosine kernel: $K(s) = \pi \cos(\pi s/2)I(|s| \leq 1)/4$.

For estimating any covariance matrix $\mathbf{\Sigma}(u_0)$ with the label $u_0 \in [0, 1]$, the smoothed sample covariance matrix estimator $\mathbf{S}(u_0)$ is calculated as follows:

$$\mathbf{S}(u_0) := \sum_{i=1}^{n} \omega_i(u_0, h)\hat{\mathbf{\Sigma}}_i, \tag{2.3}$$

where $\omega_i(u_0, h)$ is a weight function and $\hat{\mathbf{\Sigma}}_i$ is the sample covariance matrix of $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}$:

$$\omega_i(u_0, h) := \frac{c(u_0)}{nh} K\left(\frac{u_i - u_0}{h}\right), \quad \hat{\mathbf{\Sigma}}_i := \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{it}\boldsymbol{x}_{it}^{\mathsf{T}} \in \mathbb{R}^{d\times d}. \tag{2.4}$$

Here $c(u_0) = 2I(u_0 \in \{0, 1\}) + I\{u_0 \in (0, 1)\}$ is a constant depending on whether u_0 is on the boundary or not, and h is the bandwidth parameter. We will discuss how to select h in the next section.

After obtaining the covariance matrix estimate, $\mathbf{S}(u_0)$, we proceed to estimate $\mathbf{\Omega}(u_0) := \{\mathbf{\Sigma}(u_0)\}^{-1}$. When a suitable sparsity assumption on the inverse covariance matrix $\mathbf{\Omega}(u_0)$ is available, we propose to estimate $\mathbf{\Omega}(u_0)$ by plugging $\mathbf{S}(u_0)$ into the CLIME algorithm (Cai et al., 2011). In detail, the inverse covariance matrix estimator $\hat{\mathbf{\Omega}}(u_0)$ of $\mathbf{\Omega}(u_0)$ is calculated via solving the following optimization problem:

$$\hat{\mathbf{\Omega}}(u_0) = \underset{\mathbf{M}\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \sum_{jk} |\mathbf{M}_{jk}|, \quad \text{subject to } \|\mathbf{S}(u_0)\mathbf{M} - \mathbf{I}_d\|_{\max} \le \lambda, \tag{2.5}$$

where $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix and λ is a tuning parameter. Equation (2.5) can be further decomposed into d optimization subproblems (Cai et al., 2011). For $j = 1, \ldots, d$, the j-th column of $\hat{\mathbf{\Omega}}(u_0)$ can be solved as:

$$\{\hat{\mathbf{\Omega}}(u_0)\}_{*j} = \underset{\boldsymbol{v}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{v}\|_1, \quad \text{subject to } \|\mathbf{S}(u_0)\boldsymbol{v} - \boldsymbol{e}_j\|_\infty \le \lambda, \tag{2.6}$$

where $\boldsymbol{e}_j$ is the j-th canonical vector. Equation (2.6) can be solved efficiently using a

parametric simplex algorithm (Pang et al., 2013). Hence, the solution to Equation (2.5) can be computed in parallel.

Once $\hat{\mathbf{\Omega}}(u_0)$ is obtained, we can apply an additional threshold step to estimate the Graph $\mathbf{G}(u_0)$. We define a graph estimator $\hat{\mathbf{G}} \in \{0,1\}^{d\times d}$ to be:

$$\left\{\hat{\mathbf{G}}(u_0)\right\}_{jk} = \begin{cases} 1 & \text{if } \left|\left\{\hat{\mathbf{\Omega}}(u_0)\right\}_{jk}\right| > \gamma, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

Here γ is another tuning parameter.

Of note, although two tuning parameters, λ and γ, are introduced, γ is introduced merely for theoretical soundness. Empirically, we found that setting γ to be 0 or a very small value (e.g., 10^{-5}) has proven to work well. This is consistent with existing literature on graphical model estimation. We refer the readers to Cai et al. (2011), Liu et al. (2012a), Liu et al. (2012b), Xue and Zou (2012), and Han et al. (2013) for more discussion on this issue.

Procedures for choosing λ have also been well studied in the graphical model literature. On one hand, popular selection criteria, such as the stability approach based on subsampling (Meinshausen and Bühlmann, 2010; Liu et al., 2010), exist and have been well studied. On the other hand, when prior knowledge about the sparsity of the precision matrix is available, a common approach is trying a sequence of λ, and choosing one according to a desired sparsity level.

2.3 Theoretical Properties

In this section the theoretical properties of the proposed estimators in Equations (2.5) and (2.7) are provided. Under a double asymptotic framework, the rates of convergence in parameter estimation under the matrix ℓ_1 and $\ell_{\max}$ norms are given.

Before establishing the theoretical result, we first pose an additional assumption on the function $\mathbf{\Sigma}(\cdot)$. In detail, let $\Sigma_{jk}(\cdot) : u \to \{\mathbf{\Sigma}(u)\}_{jk}$ be a real function. In the following, we assume that $\Sigma_{jk}(\cdot)$ is a smooth function with regard to any $j, k \in \{1, \ldots, d\}$. Here and in the sequel, the derivatives at support boundaries are defined as one-sided derivatives.

(A1) There exists one absolute constant, C_2, such that for all $u \in [0, 1]$,

$$\left| \frac{d}{du} \Sigma_{jk}(u) \right| \leq C_2, \quad \text{for } j, k \in \{1, \ldots, d\}.$$

Under Assumption **(A1)**, we propose the following lemma, which shows that when the subjects are sampled in $u = u_1, \ldots, u_n$ with $u_i = i/n$ for $i = 1, \ldots, n$, the estimator $\mathbf{S}(u_0)$ approximates $\mathbf{\Sigma}(u_0)$ at a fast rate for any $u_0 \in [0, 1]$. The convergence rate delivered here characterizes both the strength one can borrow across different subjects and the impact of temporal dependence structure on estimation accuracy.

Lemma 1. *Suppose that the data points are generated from the model discussed in Section 2.2.1 and Assumption* **(A1)** *holds. Moreover, suppose that the observed subjects are in*

$u_i = i/n$ *for* $i = 1, \ldots, n$. *Then, for any* $u_0 \in [0, 1]$, *if for some* $\eta > 0$ *we have*

$$(\mathbf{A2}) \quad \sup_{u\in[0,1]} \frac{d^2}{du^2} \left\{ K\left(\frac{u-u_0}{h}\right) \Sigma_{jk}(u) \right\} = O(h^{-\eta}), \quad \text{for } j, k \in \{1, \ldots, d\},$$

and the bandwidth h *is set as*

$$h \asymp \max\left\{ \left\{ \frac{\xi \cdot \sup_{u\in[0,1]} \|\mathbf{\Sigma}(u)\|_2}{1 - \sup_{u\in[0,1]} \|\mathbf{A}(u)\|_2} \sqrt{\frac{\log d}{Tn}} \right\}^{1/2}, n^{-\frac{2}{2+\eta}} \right\}, \tag{2.8}$$

where $\xi := \sup_{u\in[0,1]} \max_j [\mathbf{\Sigma}(u)]_{jj} / \min_j [\mathbf{\Sigma}(u)]_{jj}$, *then the smoothed sample covariance matrix estimator* $\mathbf{S}(u_0)$ *defined in Equation* (2.3) *satisfies:*

$$\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left[\left\{ \frac{\xi \sup_{u\in[0,1]} \|\mathbf{\Sigma}(u)\|_2}{1 - \sup_{u\in[0,1]} \|\mathbf{A}(u)\|_2} \sqrt{\frac{\log d}{Tn}} \right\}^{1/2} + n^{-\frac{2}{2+\eta}} \right]. \tag{2.9}$$

Assumption **(A2)** is a convolution between the smoothness of $K(\cdot)$ and $\Sigma_{jk}(\cdot)$, and is a weaker requirement than imposing smoothness individually. Assumption **(A2)** is satisfied by many commonly used kernel functions, including the aforementioned examples in Section 2.2.2. For example, with regard to the Epanechnikov kernel $K(s) = 3(1 - s^2)I(|s| \leq 1)/4$, it's easy to check that

$$\frac{d}{du} K\left(\frac{u-u_0}{h}\right) = O\left(\frac{1}{h^2}\right) \quad \text{and} \quad \frac{d^2}{du^2} K\left(\frac{u-u_0}{h}\right) = O\left(\frac{1}{h^2}\right).$$

Therefore, as long as $\Sigma_{jk}(u)$, $\frac{d}{du}\Sigma_{jk}(u)$, and $\frac{d^2}{du^2}\Sigma_{jk}(u)$ are uniformly bounded, the Epanech-

nikov kernel satisfies Assumption **(A2)** with $\eta \geq 2$.

There are several observations drawn from Lemma 1. First, the rate of convergence in parameter estimation is upper bounded by $n^{-\frac{2}{2+\eta}}$, which is due to the bias in estimating $\mathbf{\Sigma}(u_0)$ from only n labels. This term is irrelevant to the sample size T in each subject and cannot be improved without adding stronger (potentially unrealistic) assumptions. For example, when none of ξ, $\sup_t \|\mathbf{\Sigma}(u)\|_2$, and $\sup_t \|\mathbf{A}(u)\|_2$ scales with (n, T, d) and $T > Cn^{\frac{6-\eta}{2+\eta}} \log d$ for some generic constant C, the estimator achieves a $n^{-\frac{2}{2+\eta}}$ rate of convergence. Secondly, in the term $\{\log d/(Tn)\}^{1/4}$, n characterizes the strength one can borrow across different subjects, while T demonstrates the contribution from within a subject. When $n > CT^{\frac{2+\eta}{6-\eta}}$, the estimator achieves a $\{\log d/(Tn)\}^{1/4}$ rate of convergence. The first two points discussed above, together, quantify the settings where the proposed methods can beat the naive method which only exploits the data points in each subject itself for parameter estimation.

Finally, Lemma 1 also demonstrates how temporal dependence may affect the rate of convergence. Specifically, the spectral norm of the transition matrix, $\|\mathbf{A}(u)\|_2$, characterizes the strength of temporal dependence. The term $1/\{1 - \sup_{u\in[0,1]} \|\mathbf{A}(u)\|_2\}$ in Equation (2.9) demonstrates the impact of the dependence strength on the rate of convergence. Further discussions on the effect of $\mathbf{A}(u)$ are collected in Section A.1 of the appendix.

Next, we consider the case where $\mathbf{A}(u) = 0$ and hence $\{\boldsymbol{x}_{it}\}_{t=1}^T$ are independent observations with no temporal dependence. In this case, following Zhou et al. (2010), the rate of convergence in parameter estimation can be improved.

Lemma 2. *Under the assumptions in Lemma 1, if it is further assumed that*

(B1) $\{\boldsymbol{x}_{it}\}_{t=1}^{T}$ *are i.i.d. observations from* $N_d\{\mathbf{0}, \boldsymbol{\Sigma}(u)\}$*;*

(B2) $\sup_{u\in[0,1]} \frac{d^2}{du^2}\left[K^2\left(\frac{u-u_0}{h}\right)\left\{\Sigma_{jj}^2(u)\Sigma_{kk}^2(u) + \Sigma_{jk}^2(u)\right\}\right] = O(h^{-4})$ for all $j, k \in \{1, \ldots, d\}$*;*

(B3) *There exists an absolute constant* C_3 *such that*

$$\max_{jk} \sup_{u\in[0,1]} |\Sigma_{jk}(u)| \leq C_3, \quad \max_{jk} \sup_{u\in[0,1]} \left|\frac{d}{du}\Sigma_{jk}(u)\right| \leq C_3;$$

then, setting the bandwidth

$$h \asymp \max\left\{\left(\frac{\log d}{Tn}\right)^{1/3}, \frac{1}{n^{2/(2+\eta)}}\right\}, \tag{2.10}$$

we have

$$\|\mathbf{S}(u_0) - \boldsymbol{\Sigma}(u_0)\|_{\max} = O_P\left\{\left(\frac{\log d}{Tn}\right)^{1/3} + n^{-\frac{2}{2+\eta}}\right\}.$$

We note again that the aforementioned kernel functions satisfy Assumptions **(B2)** for similar reasons. In detail, taking Epanechnikov kernel as an example, we have

$$\frac{d}{du}K^2\left(\frac{u-u_0}{h}\right) = O\left(\frac{1}{h^4}\right), \quad \frac{d^2}{du^2}K^2\left(\frac{u-u_0}{h}\right) = O\left(\frac{1}{h^4}\right).$$

So Assumption **(B2)** is satisfied as long as $\Sigma_{jk}(u)$, $\frac{d}{du}\Sigma_{jk}(u)$, and $\frac{d^2}{du^2}\Sigma_{jk}(u)$ are uniformly bounded.

Lemma 2 shows that the rate of convergence can be improved to $\{\log d/(Tn)\}^{1/3}$ when the data are independent. Of note, this rate matches the results in Zhou et al. (2010). However, the improved rate is valid only when a strong independence assumption holds, which is unrealistic in many applications, rs-fMRI data analysis for example.

After obtaining Lemmas 1 and 2, we proceed to the final result, which shows the theoretical performance of the estimators $\hat{\mathbf{\Omega}}(u_0)$ and $\hat{\mathbf{G}}(u_0)$ proposed in Equations (2.5) and (2.7). We show that under certain sparsity constraints, the proposed estimators are consistent, even when d is nearly exponentially larger than n and T.

We first introduce some additional notation. Let $M_d \in \mathbb{R}$ be a quantity which may scale with (n, T, d). We define the set of positive definite matrices in $\mathbb{R}^{d\times d}$, denoted by $\mathcal{M}(q, s, M_d)$, as

$$\mathcal{M}(q, s, M_d) := \left\{ \mathbf{M} \in \mathbb{R}^{d\times d} : \max_{1\le k\le d} \sum_{j=1}^{d} |M_{jk}|^q \le s, \|\mathbf{M}\|_1 \le M_d \right\}.$$

For $q = 0$, the class $\mathcal{M}(0, s, M_d)$ contains all the matrices with the number of nonzero entries in each column less than s and bounded ℓ_1 norm. We then let

$$\kappa(n, T, d) := \left\{ \frac{\xi \sup_{u\in[0,1]} \|\mathbf{\Sigma}(u)\|_2}{1 - \sup_{u\in[0,1]} \|\mathbf{A}(u)\|_2} \sqrt{\frac{\log d}{Tn}} \right\}^{1/2} + n^{-\frac{2}{2+\eta}}, \tag{2.11}$$

$$\kappa^*(n, T, d) := \left(\frac{\log d}{Tn}\right)^{1/3} + n^{-\frac{2}{2+\eta}}. \tag{2.12}$$

Theorem 1 presents the parameter estimation and graph estimation consistency results for

the estimators defined in Equations (2.5) and (2.7).

Theorem 1. *Suppose that the conditions in Lemma 1 hold. Assume that* $\mathbf{\Theta}(u_0) := \{\mathbf{\Sigma}(u_0)\}^{-1} \in \mathcal{M}(q, s, M_d)$ *with* $0 \leq q < 1$. *Let* $\hat{\mathbf{\Theta}}(u_0)$ *be defined in Equation* (2.5). *Then there exists a constant* C_3 *only depending on* q, *such that, whenever the tuning parameter*

$$\lambda = C_3 M_d \kappa(n, T, d)$$

is chosen, one has that

$$\|\hat{\mathbf{\Theta}}(u_0) - \mathbf{\Theta}(u_0)\|_2 = O_P\left\{M_d^{2-2q} s \kappa(n, T, d)^{1-q}\right\}.$$

Moreover, let $\hat{\mathbf{G}}(u_0)$ *be the graph estimator defined in Equation* (2.7) *with the second step tuning parameter* $\gamma = 4M_d\lambda$. *If it is further assumed that* $\mathbf{\Theta}(u_0) \in \mathcal{M}(0, s, M_d)$ *and*

$$\min_{\{j,k:|\{\mathbf{\Theta}(u_0)\}_{jk}|\neq 0\}} |\{\mathbf{\Theta}(u_0)\}_{jk}| \geq 2\gamma,$$

then

$$\mathbb{P}\left\{\hat{\mathbf{G}}(u_0) = \mathbf{G}(u_0)\right\} = 1 - o(1).$$

If the conditions in Lemma 2 hold, the above results are true with κ *replaced by* κ^*.

Theorem 1 shows that the proposed method is theoretically guaranteed to be consistent in both parameter estimation and model selection, even when the dimension d is nearly

exponentially larger than nT. Theorem 1 can be proved by following the proofs of Theorem 1 and Theorem 7 in Cai et al. (2011) and the proof is accordingly omitted.

2.4 Experiments

In this section, the empirical performance of the proposed method is investigated. This section consists of two parts. In the first, we demonstrate the performance using synthetic data, where the true generating models are known. On one hand, the proposed kernel based method is compared to several existing methods. The advantage of this new method is shown in both parameter estimation and model selection. On the other hand, implications of the theoretical results are also empirically verified. In the second part, the proposed method is applied to a large scale rs-fMRI data (the ADHD-200 data) and some potentially scientifically interesting results are explored. Additional experimental results are provided in Section A.2 of the appendix.

2.4.1 Synthetic Data

The performance of the proposed kernel-smoothing estimator (denoted as KSE) is compared to three existing methods: a naive estimator (donated as naive; details follow below), Danaher et al. (2014)'s group graphical lasso (denoted as GGL), and Guo et al. (2011)'s estimator (denoted as Guo). Throughout the simulation studies, it is assumed that the graphs are evolving from $u = 0$ to $u = 1$ continuously. Although there is one graphical model

corresponding to each $u \in [0,1]$, it is assumed that data are observed at n equally spaced points $u = 0, 1/(n-1), 2/(n-1), \ldots, 1$. For each $u = 0, 1/(n-1), 2/(n-1), \ldots, 1$, T observations were generated from the corresponding graph under a stationary VAR(1) model discussed in Equation (2.1). To generate the transition matrix, $\mathbf{A}$, the precision matrix was obtained using the R package *Huge* (Zhao et al., 2012) with graph structure "random". Then it is divided by twice its largest eigenvalue to obtain $\mathbf{A}$, so that $\|\mathbf{A}\|_2 = 0.5$. The same transition matrix is used under every label u. Our main target is to estimate the graph at $u_0 = 0$, as the endpoints represent the most difficult point for estimation. We also investigate one setting where the target label is $u_0 = 1/2$, to demonstrate the performance at a non-extreme target label.

In the following, three existing methods for comparison are reviewed. naive is obtained by first calculating the sample covariance matrix at target label u_0 using only the T observations under this label, and then plugged into the CLIME algorithm. Compared to KSE, GGL and Guo do not assume that there exists a smooth change among the graphs. Instead, they assume that the data come from n categories. That is, there are n corresponding underlying graphs that potentially share common edges, and observations are available within each category. Moreover, they assume that the observations are independent both between and within different categories. With regard to implementation, they solve the following optimization problem:

$$\max_{\mathbf{\Omega}^{(0)},\ldots,\mathbf{\Omega}^{(n)} \succ 0} \sum_{i=0}^{n} T \left\{ \log\det \mathbf{\Omega}^{(i)} - \text{trace}\left(\hat{\mathbf{\Sigma}}_i \mathbf{\Omega}^{(i)}\right) \right\} - P\left(\mathbf{\Omega}^{(0)}, \ldots, \mathbf{\Omega}^{(n)}\right),$$

where $\hat{\mathbf{\Sigma}}_i$ is the sample covariance matrix calculated based on the data under label u_i. GGL uses penalty

$$P\left(\mathbf{\Omega}^{(0)}, \ldots, \mathbf{\Omega}^{(n)}\right) = \lambda_1 \sum_{i=0}^{n} \sum_{j \neq k} |\{\mathbf{\Omega}^{(i)}\}_{jk}| + \lambda_2 \sum_{j \neq k} \sqrt{\sum_{i=0}^{n} \{\mathbf{\Omega}^{(i)}\}_{jk}^2},$$

and Guo uses penalty

$$P\left(\mathbf{\Omega}^{(0)}, \ldots, \mathbf{\Omega}^{(n)}\right) = \lambda \sum_{j \neq k} \sqrt{\sum_{i=0}^{n} |\{\mathbf{\Omega}^{(i)}\}_{jk}|}.$$

Here the regularity coefficients λ_1, λ_2, and λ control the sparsity level. Danaher et al. (2014) also proposed the fused graphical lasso that separately controls sparsity of and similarity between the graphs. However, this method is not scalable when the number of categories is large and therefore not included in our comparison.

After obtaining the estimated graph, $\hat{\mathbf{G}}(u_0)$, of the true traph $\mathbf{G}(u_0)$, the model selection performance is further investigated by comparing the ROC curves of the four competing methods. Let $\hat{\mathrm{E}}(u_0)$ be the set of estimated edges corresponding to $\hat{\mathbf{G}}(u_0)$, and $\mathrm{E}(u_0)$ the set of true edges corresponding to $\mathbf{G}(u_0)$. The true positive rate (TPR) and false positive rate (FPR) are defined as

$$\text{TPR} = \frac{|\hat{\mathrm{E}}(u_0) \bigcap \mathrm{E}(u_0)|}{|\mathrm{E}(u_0)|}, \quad \text{FPR} = \frac{|\hat{\mathrm{E}}(u_0) \setminus \mathrm{E}(u_0)|}{d(d-1)/2 - |\mathrm{E}(u_0)|},$$

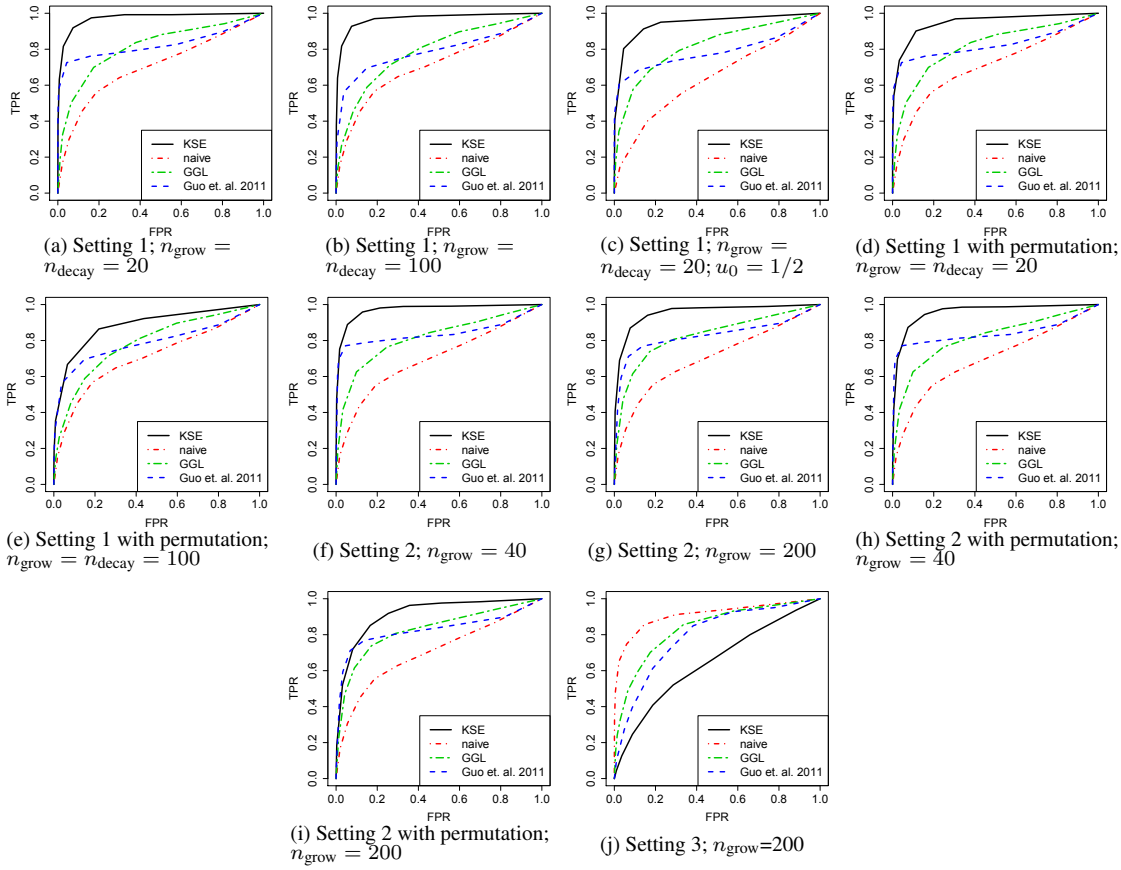
where for any set S, $|\mathrm{S}|$ denotes the cardinality of S. To obtain a series of TPRs and FPRs,

for KSE, naive, and Guo, the values of λ are varied. For GGL, first λ_2 is fixed and subsequently λ_1 is tuned, and then the λ_2 with the best overall performance is selected. More specifically, a series of λ_2 are picked, and for each fixed λ_2, λ_1 is accordingly varied to produce an ROC curve. Of note, in the investigation, the ROC curves indexed by λ_2 are generally parallel, thus motivating this strategy. Finally, the λ_2 corresponding to the topleft most curve is selected.

2.4.1.1 Setting 1: Simultaneously Evolving Edges

In this section we investigate the performance of the four competing methods under one particular graphical model. In each simulation, $n_{\text{fix}} = 200$ edges are randomly selected from $d(d-1)/2$ potential edges and they do not change with regard to the label u. The strengths of these edges, i.e. the corresponding entries in the inverse covariance matrix, are generated from a uniform distribution taking values in $[-0.3, -0.1]$ (denoted by Unif$[-0.3, -0.1]$) and do not change with u. We then randomly select n_{decay} and n_{grow} edges that will disappear and emerge over the evolution simultaneously. For each of the n_{decay} edges, the strength is generated from Unif[-0.3,-0.1] at $u = 0$ and will diminish to 0 linearly with regard to u. For each of the n_{grow} edges, the strength is set to be 0 at $u = 0$, and will linearly grow to a value generated from Unif[-0.3,-0.1]. The edges evolve simultaneously. For $j \neq k$, when we subtract a value a from $\mathbf{\Omega}_{jk}$ and $\mathbf{\Omega}_{kj}$, we increase $\mathbf{\Omega}_{jj}$ and $\mathbf{\Omega}_{kk}$ by a, and then further add 0.25 to the diagonal of the matrix to keep it positive definite.

The ROC curves under this setting with different values of n_{grow} and n_{decay} are shown



(a) Setting 1; $n_{\text{grow}} = n_{\text{decay}} = 20$

(b) Setting 1; $n_{\text{grow}} = n_{\text{decay}} = 100$

(c) Setting 1; $n_{\text{grow}} = n_{\text{decay}} = 20$; $u_0 = 1/2$

(d) Setting 1 with permutation; $n_{\text{grow}} = n_{\text{decay}} = 20$

(e) Setting 1 with permutation; $n_{\text{grow}} = n_{\text{decay}} = 100$

(f) Setting 2; $n_{\text{grow}} = 40$

(g) Setting 2; $n_{\text{grow}} = 200$

(h) Setting 2 with permutation; $n_{\text{grow}} = 40$

(i) Setting 2 with permutation; $n_{\text{grow}} = 200$

(j) Setting 3; n_{grow}=200

Figure 2.1: ROC curves of four competing methods under three settings: simultaneous (a-e), sequential (f-i), and random (j). The target labels are $u_0 = 0$ except for in (c), where $u_0 = 1/2$. In each setting we set the dimension $d = 50$, the number of labels $n = 51$, the number of observations $T = 100$, and the result is obtained by 1,000 simulations.

in Figures 2.1(a) and 2.1(b). We fix the number of labels $n = 51$, number of observations under each label $T = 100$, and dimension $d = 50$. The target label is $u_0 = 0$. It can be observed that, under both cases, KSE outperforms the other three competing methods. Moreover, when we increase the values of n_{grow} and n_{decay} from 20 to 100, the ROC curve of KSE hardly changes, since the degree of smoothness in graphical model evolving hardly change. In contrast, the ROC curves of GGL and Guo drop, since the degree of similar-

ity among the graphs is reduced. Finally, naive performances worst, which is expected because it does not borrow strength across labels in estimation. Figure 2.1(c) illustrates the performance under the same setting as in Figure 2.1(a) except $u_0 = 1/2$. KSE still outperforms the other estimators.

Next, we exploit the same data, but permute the labels $u = 1/50, 2/50, \ldots, 1$ so that the evolving pattern is much more opaque. Figures 2.1(d) and 2.1(e) illustrate the model selection result. We observe that under this setting, the ROC curves of the proposed method drop a little bit, but is still higher than the competing approaches. This is because the proposed method still benefits from the evolving graph structure (although more turbulent this time). The improvement over the naive method demonstrates exactly the strength borrowed across different labels. Note that the ROC curves of GGL, naive, and Guo shown in Figures 2.1(d) and 2.1(e) do not change compared to those in Figures 2.1(a) and 2.1(b), respectively, because they do not assume any ordering between the graphs.

2.4.1.2 Setting 2: Sequentially Growing Edges

Setting 2 is similar to Setting 1. The two differences are: (i) Here n_{decay} is set to be zero; (ii) The n_{grow} edges emerges sequentially instead of simultaneously. These n_{grow} edges are randomly selected, but there is no overlap with the existing 200 pre-fixed edges. The entries of the inverse covariance matrix for the n_{grow} edges each grow to a value generated from Unif$[-0.3, -0.1]$, linearly in a length $1/n_{\text{grow}}$ interval in $[0, 1]$, one after another. We note that there is possibility that $n < n_{\text{grow}}$, because n represents only the number of

labels we observe. Under this setting, Figures 2.1(f) and 2.1(g) plot the ROC curves of the four competing methods. We also apply the four methods to the setting where the same permutation as in Setting 1 is exploited. We show the results in Figures 2.1(h) and 2.1(i). Here the same observations persist as in Setting 1.

2.4.1.3 Setting 3: Random Edges

In this setting, in contrast to the above two settings, we violate the smoothness assumption of KSE to the extreme. We demonstrate the limitedness of the proposed method in this setting. More specifically, in this setting, under every label u, n_{ed} edges are random selected with strengths from Unif$[-0.3, -0.1]$. In this case, the graphs do not evolve smoothly over the label u, and the data under the labels $u \neq 0$ only contribute noises. We then apply the four competing methods to this setting and Figure 2.1(j) illustrates the result. Under this setting, we observe that naive beats all the other three methods. It is expected because naive is the only method that do not suffer from the noises. Here KSE performs worse than GGL and Guo, because there does not exist a natural ordering among the graphs.

Under the above three data generating settings, we further quantitatively compare the performance in parameter estimation of the inverse covariance matrix $\mathbf{\Omega}(u_0)$ for the four competing methods. Here the distances between the estimated and the true concentration matrices with regard to the matrix ℓ_1, ℓ_2, and Frobenius norms are shown in Table 2.1. It can be observed that KSE achieves the lowest estimation error in all settings except for the

Table 2.1: Comparison of inverse covariance matrix estimation errors in three data generating models. The parameter estimation error with regard to the matrix ℓ_1, ℓ_2, and Frobenius norms (denoted as ℓ_F here) is provided with standard deviations in parentheses. The results are obtained by 1,000 simulations.

		KSE			naive		
	$n_{\text{grow}} = n_{\text{decay}}$	ℓ_1	ℓ_2	ℓ_F	ℓ_1	ℓ_2	ℓ_F
Setting 1	20	3.25(0.232)	1.53(0.104)	4.42(0.220)	5.02(0.287)	2.68(0.132)	8.30(0.412)
	100	2.72(0.165)	1.30(0.088)	3.78(0.204)	4.85(0.467)	2.55(0.117)	8.13(0.453)
	n_{grow}						
Setting 2	40	3.39(0.553)	1.56(0.213)	4.47(0.302)	5.26(0.740)	2.73(0.313)	8.24(0.386)
	200	3.40(0.507)	1.57(0.147)	4.33(0.284)	5.19(0.740)	2.71(0.280)	8.34(0.352)
Setting 3	n_{ed}						
	50	2.21(0.194)	1.37(0.120)	3.20(0.104)	1.60(0.249)	0.84(0.113)	3.09(0.185)
		GGL			Guo		
	n_{grow}=n_{decay}	ℓ_1	ℓ_2	ℓ_F	ℓ_1	ℓ_2	ℓ_F
Setting 1	20	3.28(0.298)	1.45(0.112)	4.13(0.190)	3.22(0.418)	1.42(0.259)	4.04(0.280)
	100	3.27(0.324)	1.42(0.100)	4.18(0.222)	3.38(0.474)	1.41(0.169)	4.31(0.335)
	n_{grow}						
Setting 2	40	3.47(0.580)	1.47(0.163)	4.22(0.153)	3.06(0.417)	1.40(0.274)	4.00(0.205)
	200	3.22(0.618)	1.44(0.198)	4.08(0.199)	3.71(0.493)	1.73(0.264)	4.46(0.361)
Setting 3	n_{ed}						
	50	1.52(0.224)	0.85(0.105)	2.04(0.104)	1.48(0.263)	0.67(0.116)	1.81(0.150)

Setting 3. This coincides with the above model selection results. We omit the results for the label permutation cases and the case with $u_0 = 1/2$, since they are again as expected from the model selection results above.

2.4.2 Impact of a Small Label Size n

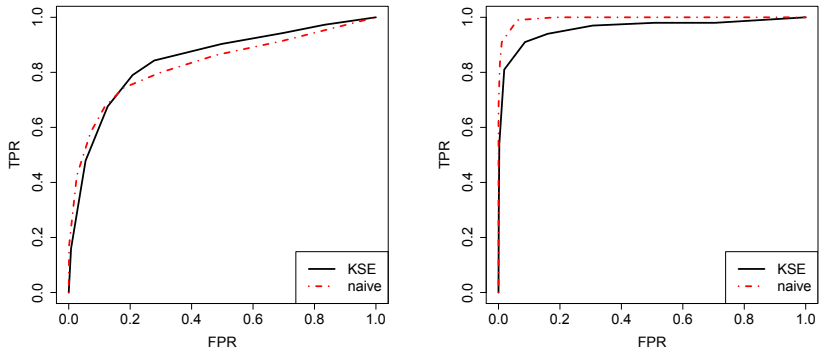
As is shown in Lemma 1 and Theorem 1, the rates of convergence in parameter estimation and model selection crucially depend on the term $n^{-\frac{2}{2+\eta}}$. This is due to the bias

in estimating $\mathbf{\Sigma}(u_0)$ from n labels. This bias takes place as long as we include data under other labels into estimation, and cannot be removed by simply increasing the number of observations T under each label u. More specifically, Lemma A.3.1 of the appendix shows that the rate of convergence for bias between the estimated and the true covariance matrix depends on n but not T.

This section is devoted to illustrate this phenomenon empirically. We exploit Setting 2 in the last section with the number of labels n to be very small. Here we set $n = 3$. Moreover, we choose $n_{\text{fix}} = 100$, $n_{\text{grow}} = 500$, and vary the number of observations T under each label. Figure 2.2 compares the ROC curves of KSE and naive corresponding to the settings when $T = 100$ or 500. There are two important observations we would like to emphasize: (i) When $T = 100$, KSE and naive have comparable performance. However, when $T = 500$, naive performs much better than KSE. (ii) The change of the ROC curves for KSE from $T = 100$ to $T = 500$ is less dramatic compared to the ROC curves for naive. These observations indicate the existence of bias in KSE that cannot be eliminated by only increasing T.

2.4.3 ADHD-200 Data

As an example of real data application, we apply the proposed method to the ADHD-200 data (Biswal et al., 2010). The ADHD-200 data consist of rs-fMRI images of 973 subjects. Of them, 491 are healthy and 197 have been diagnosed with ADHD type 1,2, or 3. The remaining had their diagnosis withheld for the purpose of a prediction competition.



(a) $T = 100$ (b) $T = 500$

Figure 2.2: ROC curves of KSE and naive under Setting 1: sequentially evolving edges. We set dimension $d = 50$; number of labels $n = 3$; number of pre-fixed edges $n_{\text{fix=100}}$; number of growing edges $n_{\text{grow}} = 500$.

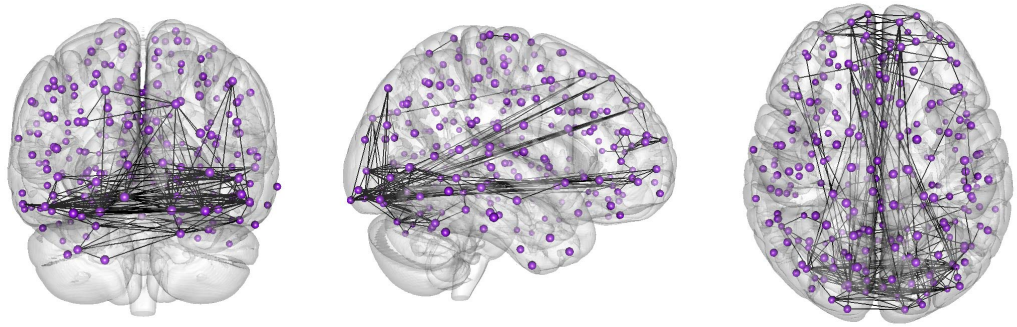
The number of images for each subject ranges from 76 to 276. 264 seed regions of interest are used to define nodes for graphical model analysis (Power et al., 2011). A limited set of covariates including gender, age, handedness, IQ, are available.

2.4.3.1 Brain Development

In this section, focus lies on investigating the development of brain connectivity network over age for control subjects. Here the subject ages are normalized to be in $[0, 1]$, and the brain ROI measurements are centered to have sample means zero and scaled to have sample standard deviations 1. The bandwidth parameter is set at $h = 0.5$. The regularization parameter λ is manually chosen to induce high sparsity for better visualization and highlighting the dominating edges. Consider estimating the brain networks at ages 7.09, 11.75, and 21.83, which are the minimal, median, and maximal ages in the data. Figure 2.3 shows coronal, sagittal, and transverse snapshots of the estimated brain connectivity

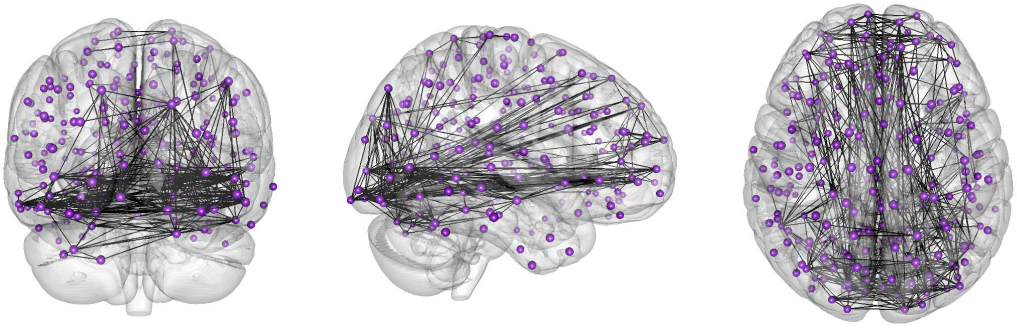
networks.

There are two main patterns worth noting in this experiment: (i) It is observed that the degree of complexity of the brain network at the occipital lobe is high compared to other regions by age seven. This is consistent with early maturation of visual and vision processing networks relative to others. We found that this conjecture is supported by several recent scientific results (Shaw et al., 2008; Blakemore, 2012). For example, Shaw et al. (2008) showed that occipital lobe is fully developed before other brain regions. Moreover, when considering structural development, the occipital lobe reaches its peak thickness by age nine. In comparison, portions of the parietal lob reaches their peak thickness as late as thirteen. (ii) Figure 2.3 also shows that dense connections in the temporal lobe only occur in the graph at age 21.83 among the ages shown. This is also supported by the scientific finding that grey matter in the temporal lobe doesn't reach maximum volume untill age 16 (Bartzokis et al., 2001; Giedd et al., 1999). We also noticed that several confounding factors, such as scanner noise, subject motion, and coregistration, can have potential effects on inference (Braun et al., 2012; Van Dijk et al., 2012). In this manuscript, we rely on the standard data pre-processing techniques as described in Eloyan et al. (2012) for removing such confounders. The influence of these confounders on our inference will be investigated in greater detail in the future.



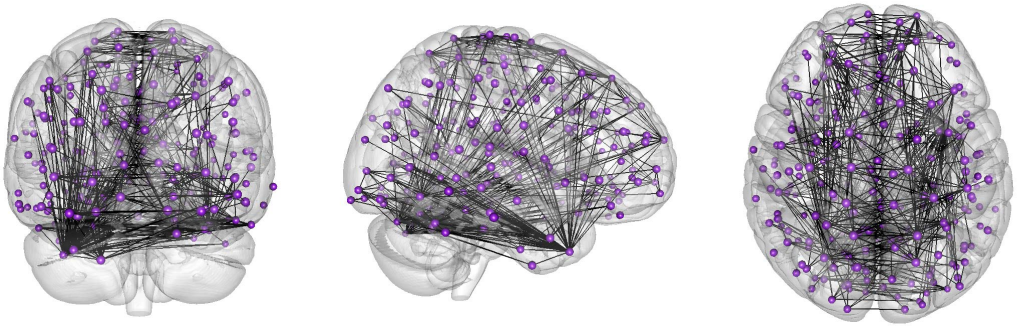
(a) coronal (b) sagittal (c) transverse

Age 7.09



(d) coronal (e) sagittal (f) transverse

Age 11.75



(g) coronal (h) sagittal (i) transverse

Age 21.83

Figure 2.3: Estimated brain connectivity network at ages 7.09, 11.75, 21.83 in healthy subjects.

2.5 Discussion

In this paper, we introduced a new kernel based estimator for jointly estimating multiple graphical models under the condition that the models smoothly vary according to a label. Methodologically, motivated by resting state functional brain connectivity analysis, we proposed a new model, taking both heterogeneity structure and dependence issues into consideration, and introduced a new kernel based method under this model. Theoretically, we provided the model selection and parameter estimation consistency result for the proposed method under both the independence and dependence assumptions. Empirically, we applied the proposed method to synthetic and real brain image data. We found that the proposed method is effective for both parameter estimation and model selection compared to several existing methods under various settings.

Acknowledgement

We would like to thank John Muschelli for providing the R tools to visualize the brain network. We would also like to thank one anonymous referee, the associate editor, and the editor of Journal of the Royal Statistical Society for their helpful comments and suggestions. In addition, thanks also to Drs. Mladen Kolar, Derek Cummings, Martin Lindquist, Michelle Carlson, and Daniel Robinson for helpful discussions on this work.

Chapter 3

Robust Portfolio Optimization

3.1 Introduction

Markowitz's mean-variance analysis sets the basis for modern portfolio optimization theory (Markowitz, 1952). However, the mean-variance analysis has been criticized for being sensitive to estimation errors in the mean and covariance matrix of the asset returns (Best and Grauer, 1991; Chopra and Ziemba, 1993). Compared to the covariance matrix, the mean of the asset returns is more influential and harder to estimate (Merton, 1980; Kallberg and Ziemba, 1984). Therefore, many studies focus on the global minimum variance (GMV) formulation, which only involves estimating the covariance matrix of the asset returns.

Estimating the covariance matrix of asset returns is challenging due to the high dimensionality and heavy-tailedness of asset return data. Specifically, the number of assets under management is usually much larger than the sample size of exploitable historical data. On the other hand, extreme events are typical in financial asset prices, leading to heavy-tailed asset returns.

To overcome the curse of dimensionality, structured covariance matrix estimators are proposed for asset return data. Fan et al. (2008) considered estimators based on factor models with observable factors. Stock and Watson (2002); Bai et al. (2012); Fan et al. (2013a) studied covariance matrix estimators based on latent factor models. Ledoit and Wolf (2003, 2004a,b) proposed to shrink the sample covariance matrix towards highly structured covariance matrices, including the identity matrix, order 1 autoregressive covariance matrices, and one-factor-based covariance matrix estimators. These estimators are commonly based

on the sample covariance matrix. (sub)Gaussian tail assumptions are required to guarantee consistency.

For heavy-tailed data, robust estimators of covariance matrices are desired. Classic robust covariance matrix estimators include M-estimators, minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators, S-estimators, and estimators based on data outlyingness and depth (Huber, 1981). These estimators are specifically designed for data with very low dimensions and large sample sizes. For generalizing the robust estimators to high dimensions, Maronna and Zamar (2002) proposed the Orthogonalized Gnanadesikan-Kettenring (OGK) estimator, which extends Gnanadesikan and Kettenring (1972)'s estimator by re-estimating the eigenvalues; Chen et al. (2011b); Couillet and McKay (2014) studied shrinkage estimators based on Tyler's M-estimator. However, although OGK is computationally tractable in high dimensions, consistency is only guaranteed under fixed dimension. The shrunken Tylor's M-estimator involves iteratively inverting large matrices. Moreover, its consistency is only guaranteed when the dimension is in the same order as the sample size. The aforementioned robust estimators are analyzed under independent data points. Their performance under time series data is questionable.

In this paper, we build on a quantile-based scatter matrix[1] estimator, and propose a robust portfolio optimization approach. Our contributions are in three aspects. First, we show that the proposed method accommodates high dimensional data by allowing the dimension to scale exponentially with sample size. Secondly, we verify that consistency of the pro-

[1] A scatter matrix is defined to be any matrix proportional to the covariance matrix by a constant.

posed method is achieved without any tail conditions, thus allowing for heavy-tailed asset return data. Thirdly, we consider weakly dependent time series, and demonstrate how the degree of dependence impacts the consistency of the proposed method.

3.2 Background

In this section, we introduce the notation system, and provide a review on the gross-exposure constrained portfolio optimization that will be exploited in this paper.

3.2.1 Notation

Let $\mathbf{v} = (v_1, \ldots, v_d)^\mathsf{T}$ be a d-dimensional real vector, and $\mathbf{M} = [\mathbf{M}_{jk}] \in \mathbb{R}^{d_1 \times d_2}$ be a $d_1 \times d_2$ matrix with $\mathbf{M}_{jk}$ as the (j,k) entry. For $0 < q < \infty$, we define the ℓ_q vector norm of $\mathbf{v}$ as $\|\mathbf{v}\|_q := (\sum_{j=1}^d |v_j|)^{1/q}$ and the ℓ_∞ vector norm of $\mathbf{v}$ as $\|\mathbf{v}\|_\infty := \max_{j=1}^d |v_j|$. Let the matrix $\ell_{\max}$ norm of $\mathbf{M}$ be $\|\mathbf{M}\|_{\max} := \max_{jk} |M_{jk}|$, and the Frobenius norm be $\|\mathbf{M}\|_F := \sqrt{\sum_{jk} M_{jk}^2}$. Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\mathsf{T}$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\mathsf{T}$ be two random vectors. We write $\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{Y}$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed. We use $\mathbf{1}, \mathbf{2}, \ldots$ to denote vectors with $1, 2, \ldots$ at every entry.

3.2.2 Gross-exposure Constrained Global Minimum Variance Formulation

Under the global minimum variance (GMV) formulation, Jagannathan and Ma (2003) found that imposing a no-short-sale constraint improves portfolio efficiency. Fan et al. (2012a) relaxed the no-short-sale constraint by a gross-exposure constraint, and showed that portfolio efficiency can be further improved.

Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector of asset returns. A portfolio is characterized by a vector of investment allocations, $\mathbf{w} = (w_1, \ldots, w_d)^\mathsf{T}$, among the d assets. The gross-exposure constrained GMV portfolio optimization can be formulated as

$$\min_{\mathbf{w}} \mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{w} \;\text{ s.t. }\; \mathbf{1}^\mathsf{T}\mathbf{w} = 1,\; \|\mathbf{w}\|_1 \leq c. \tag{3.1}$$

Here $\mathbf{1}^\mathsf{T}\mathbf{w} = 1$ is the budget constraint, $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$, and $\|\mathbf{w}\|_1 \leq c$ is the gross-exposure constraint. $c \geq 1$ is called the gross exposure constant, which controls the percentage of long and short positions allowed in the portfolio (Fan et al., 2012a). The optimization problem (3.1) can be converted into a quadratic programming problem, and solved by standard software (Fan et al., 2012a).

3.3 Method

In this section, we introduce the quantile-based portfolio optimization approach. Let $Z \in \mathbb{R}$ be a random variable with distribution function F, and $\{z_t\}_{t=1}^T$ be a sequence of observations from Z. For a constant $q \in [0,1]$, we define the q-quantiles of Z and $\{z_t\}_{t=1}^T$ to be

$$Q(Z;q) = Q(F;q) := \inf\{z : \mathbb{P}(Z \le z) \ge q\},$$
$$\hat{Q}(\{z_t\}_{t=1}^T; q) := z^{(k)} \text{ where } k = \min\Big\{t : \frac{t}{T} \ge q\Big\}.$$

Here $z^{(1)} \le \ldots \le z^{(T)}$ are the order statistics of $\{z_t\}_{t=1}^T$. We say $Q(Z;q)$ is unique if there exists a unique z such that $\mathbb{P}(Z \le z) = q$. We say $\hat{Q}(\{z_t\}_{t=1}^T; q)$ is unique if there exists a unique $z \in \{z_1, \ldots, z_T\}$ such that $z = z^{(k)}$. Following the estimator Q_n (Rousseeuw and Croux, 1993), we define the population and sample quantile-based scales to be

$$\sigma^{\mathrm{Q}}(Z) := Q(|Z - \tilde{Z}|; 1/4) \text{ and } \hat{\sigma}^{\mathrm{Q}}(\{z_t\}_{t=1}^T) := \hat{Q}(\{|z_s - z_t|\}_{1 \le s < t \le T}; 1/4). \tag{3.2}$$

Here $\tilde{Z}$ is an independent copy of Z. Based on σ^{Q} and $\hat{\sigma}^{\mathrm{Q}}$, we can further define robust scatter matrices for asset returns. In detail, let $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathsf{T}} \in \mathbb{R}^d$ be a random vector representing the returns of d assets, and $\{\boldsymbol{X}_t\}_{t=1}^T$ be a sequence of observations from $\boldsymbol{X}$, where $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{td})^{\mathsf{T}}$. We define the population and sample quantile-

based scatter matrices (QNE) to be

$$\mathbf{R}^{\mathrm{Q}} := [\mathbf{R}^{\mathrm{Q}}_{jk}] \text{ and } \hat{\mathbf{R}}^{\mathrm{Q}} := [\hat{\mathbf{R}}^{\mathrm{Q}}_{jk}],$$

where the entries of $\mathbf{R}^{\mathrm{Q}}$ and $\hat{\mathbf{R}}^{\mathrm{Q}}$ are given by

$$\begin{aligned}
\mathbf{R}^{\mathrm{Q}}_{jj} &:= \sigma^{\mathrm{Q}}(X_j)^2, \;\; \hat{\mathbf{R}}^{\mathrm{Q}}_{jj} := \hat{\sigma}^{\mathrm{Q}}(\{X_{tj}\}_{t=1}^{T})^2, \\
\mathbf{R}^{\mathrm{Q}}_{jk} &:= \frac{1}{4}\Big[\sigma^{\mathrm{Q}}(X_j + X_k)^2 - \sigma^{\mathrm{Q}}(X_j - X_k)^2\Big], \\
\hat{\mathbf{R}}^{\mathrm{Q}}_{jk} &:= \frac{1}{4}\Big[\hat{\sigma}^{\mathrm{Q}}(\{X_{tj} + X_{tk}\}_{t=1}^{T})^2 - \sigma^{\mathrm{Q}}(\{X_{tj} - X_{tk}\}_{t=1}^{T})^2\Big].
\end{aligned}$$

Since $\hat{\sigma}^{\mathrm{Q}}$ can be computed using $O(T \log T)$ time (Rousseeuw and Croux, 1993), the computational complexity of $\hat{\mathbf{R}}^{\mathrm{Q}}$ is $O(d^2 T \log T)$. Since $T \ll d$ in practice, $\hat{\mathbf{R}}^{\mathrm{Q}}$ can be computed almost as efficiently as the sample covariance matrix, which has $O(d^2 T)$ complexity.

Let $\mathbf{w} = (w_1, \ldots, w_d)^{\mathsf{T}}$ be the vector of investment allocations among the d assets. For a matrix $\mathbf{M}$, we define a risk function $R : \mathbb{R}^d \times \mathbb{R}^{d \times d} \to \mathbb{R}$ by

$$R(\mathbf{w}; \mathbf{M}) := \mathbf{w}^{\mathsf{T}} \mathbf{M} \mathbf{w}.$$

When $\boldsymbol{X}$ has covariance matrix $\boldsymbol{\Sigma}$, $R(\mathbf{w}; \boldsymbol{\Sigma}) = \mathrm{Var}(\mathbf{w}^{\mathsf{T}} \boldsymbol{X})$ is the variance of the portfolio return, $\mathbf{w}^{\mathsf{T}} \boldsymbol{X}$, and is employed as the objected function in the GMV formulation. However, estimating $\boldsymbol{\Sigma}$ is difficult due to the heavy tails of asset returns. In this paper, we adopt $R(\mathbf{w}; \mathbf{R}^{\mathrm{Q}})$ as a robust alternative to the moment-based risk metric, $R(\mathbf{w}; \boldsymbol{\Sigma})$, and consider

the following oracle portfolio optimization problem:

$$\mathbf{w}^{\text{opt}} = \underset{\mathbf{w}}{\text{argmin}}\, R(\mathbf{w}; \mathbf{R}^{\text{Q}}) \text{ s.t. } \mathbf{1}^{\mathsf{T}}\mathbf{w} = 1, \|\mathbf{w}\|_1 \leq c. \tag{3.3}$$

Here $\|\mathbf{w}\|_1 \leq c$ is the gross-exposure constraint introduced in Section 3.2.2. In practice, $\mathbf{R}^{\text{Q}}$ is unknown and has to be estimated. For convexity of the risk function, we project $\hat{\mathbf{R}}^{\text{Q}}$ onto the cone of positive definite matrices:

$$\begin{aligned} &\tilde{\mathbf{R}}^{\text{Q}} = \text{argmin}_{\mathbf{R}} \left\| \hat{\mathbf{R}}^{\text{Q}} - \mathbf{R} \right\|_{\max} \\ &\text{s.t. } \mathbf{R} \in S_\lambda := \{\mathbf{M} \in \mathbb{R}^{d\times d} : \mathbf{M}^{\mathsf{T}} = \mathbf{M}, \lambda_{\min}\mathbf{I}_d \preceq \mathbf{M} \preceq \lambda_{\max}\mathbf{I}_d\}. \end{aligned} \tag{3.4}$$

Here $\lambda_{\min}$ and $\lambda_{\max}$ set the lower and upper bounds for the eigenvalues of $\tilde{\mathbf{R}}^{\text{Q}}$. The optimization problem (3.4) can be solved by a projection and contraction algorithm (Xu and Shao, 2012b). We summarize the algorithm in the Appendix B.3. Using $\tilde{\mathbf{R}}^{\text{Q}}$, we formulate the empirical robust portfolio optimization by

$$\tilde{\mathbf{w}}^{\text{opt}} = \underset{\mathbf{w}}{\text{argmin}}\, R(\mathbf{w}; \tilde{\mathbf{R}}^{\text{Q}}) \text{ s.t. } \mathbf{1}^{\mathsf{T}}\mathbf{w} = 1,\ \|\mathbf{w}\|_1 \leq c. \tag{3.5}$$

Remark 2. The robust portfolio optimization approach involves three parameters: $\lambda_{\min}$, $\lambda_{\max}$, and c. Empirically, setting $\lambda_{\min} = 0.005$ and $\lambda_{\max} = \infty$ proves to work well. c is typically provided by investors for controlling the percentages of short positions. When a data-driven choice is desired, we refer to Fan et al. (2012a) for a cross-validation-based

approach.

Remark 3. The rationale behind the positive definite projection (3.4) lies in two aspects. First, in order that the portfolio optimization is convex and well conditioned, a positive definite matrix with lower bounded eigenvalues is needed. This is guaranteed by setting $\lambda_{\min} > 0$. Secondly, the projection (3.4) is more robust compared to the OGK estimate (Maronna and Zamar, 2002). OGK induces positive definiteness by re-estimating the eigenvalues using the variances of the principal components. Robustness is lost when the data, possibly containing outliers, are projected onto the principal directions for estimating the principal components.

Remark 4. We adopt the $1/4$ quantile in the definitions of σ^{Q} and $\hat{\sigma}^{\mathrm{Q}}$ to achieve 50% breakdown point. However, we note that our methodology and theory carries through if $1/4$ is replaced by any absolute constant $q \in (0, 1)$.

3.4 Theoretical Properties

In this section, we provide theoretical analysis of the proposed portfolio optimization approach. For an optimized portfolio, $\hat{\mathbf{w}}^{\mathrm{opt}}$, based on an estimate, $\mathbf{R}$, of $\mathbf{R}^{\mathrm{Q}}$, the next lemma shows that the error between the risks $R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}})$ and $R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}})$ is essentially related to the estimation error in $\mathbf{R}$.

Lemma 3. *Let* $\hat{\mathbf{w}}^{\mathrm{opt}}$ *be the solution to*

$$\min_{\mathbf{w}} R(\mathbf{w}; \mathbf{R}) \text{ s.t. } \mathbf{1}^{\mathsf{T}}\mathbf{w} = 1,\ \|\mathbf{w}\|_1 \leq c \tag{3.6}$$

for an arbitrary matrix $\mathbf{R}$*. Then, we have*

$$|R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) - R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}})| \leq 2c^2 \|\mathbf{R} - \mathbf{R}^{\mathrm{Q}}\|_{\max},$$

where $\mathbf{w}^{\mathrm{opt}}$ *is the solution to the oracle portfolio optimization problem* (3.3)*, and* c *is the gross-exposure constant.*

Next, we derive the rate of convergence for $R(\tilde{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}})$, which relates to the rate of convergence in $\|\tilde{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max}$. To this end, we first introduce a dependence condition on the asset return series.

Definition 5. *Let* $\{X_t\}_{t\in Z}$ *be a stationary process. Denote by* $\mathcal{F}_{-\infty}^{0} := \sigma(X_t : t \leq 0)$ *and* $\mathcal{F}_{n}^{\infty} := \sigma(X_t : t \geq n)$ *the* σ*-fileds generated by* $\{X_t\}_{t\leq 0}$ *and* $\{X_t\}_{t\geq n}$*, respectively. The* ϕ*-mixing coefficient is defined by*

$$\phi(n) := \sup_{B\in\mathcal{F}_{-\infty}^{0}, A\in\mathcal{F}_{n}^{\infty}, \mathbb{P}(B)>0} |\mathbb{P}(A \mid B) - \mathbb{P}(A)|.$$

The process $\{X_t\}_{t\in\mathbb{Z}}$ *is* ϕ*-mixing if and only if* $\lim_{n\to\infty}\phi(n) = 0$*.*

Condition 1. $\{\boldsymbol{X}_t \in \mathbb{R}^d\}_{t\in\mathbb{Z}}$ *is a stationary process such that for any* $j \neq k \in \{1, \ldots, d\}$*,* $\{X_{tj}\}_{t\in\mathbb{Z}}$*,* $\{X_{tj} + X_{tk}\}_{t\in\mathbb{Z}}$*, and* $\{X_{tj} - X_{tk}\}_{t\in\mathbb{Z}}$ *are* ϕ*-mixing processes satisfying* $\phi(n) \leq$

$1/n^{1+\epsilon}$ *for any* $n > 0$ *and some constant* $\epsilon > 0$.

The parameter ϵ determines the rate of decay in $\phi(n)$, and characterizes the degree of dependence in $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$. Next, we introduce an identifiability condition on the distribution function of the asset returns.

Condition 2. *Let* $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_d)^\intercal$ *be an independent copy of* $\boldsymbol{X}_1$. *For any* $j \neq k \in \{1, \ldots, d\}$, *let* $F_{1;j}$, $F_{2;j,k}$, *and* $F_{3;j,k}$ *be the distribution functions of* $|X_{1j} - \tilde{X}_j|$, $|X_{1j} + X_{1k} - \tilde{X}_j - \tilde{X}_k|$, *and* $|X_{1j} - X_{1k} - \tilde{X}_j + \tilde{X}_k|$. *We assume there exist constants* $\kappa > 0$ *and* $\eta > 0$ *such that*

$$\inf_{|y - Q(F;1/4)| \leq \kappa} \frac{d}{dy} F(y) \geq \eta$$

for any $F \in \{F_{1;j}, F_{2;j,k}, F_{3;j,k} : j \neq k = 1, \ldots, d\}$.

Condition 2 guarantees the identifiability of the $1/4$ quantiles, and is standard in the literature on quantile statistics (Belloni and Chernozhukov, 2011; Wang et al., 2012). Based on Conditions 1 and 2, we can present the rates of convergence for $\hat{\mathbf{R}}^{\mathrm{Q}}$ and $\tilde{\mathbf{R}}^{\mathrm{Q}}$.

Theorem 6. *Let* $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ *be an absolutely continuous stationary process satisfying Conditions 1 and 2. Suppose* $\log d/T \to 0$ *as* $T \to \infty$. *Then, for any* $\alpha \in (0, 1)$ *and* T *large enough, with probability no smaller than* $1 - 8\alpha^2$, *we have*

$$\|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \leq r_T. \tag{3.7}$$

Here the rate of convergence r_T is defined by

$$r_T = \max\Big\{ \frac{2}{\eta^2}\Big[\sqrt{\frac{4(1+2C_\epsilon)(\log d - \log\alpha)}{T}} + \frac{4C_\epsilon}{T}\Big]^2, \\ \frac{4\sigma^{\mathrm{Q}}_{\max}}{\eta}\Big[\sqrt{\frac{4(1+2C_\epsilon)(\log d - \log\alpha)}{T}} + \frac{4C_\epsilon}{T}\Big]\Big\}, \tag{3.8}$$

where $\sigma^{\mathrm{Q}}_{\max} := \max\{\sigma^{\mathrm{Q}}(X_j), \sigma^{\mathrm{Q}}(X_j + X_k), \sigma^{\mathrm{Q}}(X_j - X_k) : j \neq k \in \{1, \ldots, d\}\}$ *and* $C_\epsilon := \sum_{k=1}^{\infty} 1/k^{1+\epsilon}$. *Moreover, if* $\mathbf{R}^{\mathrm{Q}} \in S_\lambda$ *for* S_λ *defined in* (3.4)*, we further have*

$$\|\tilde{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \leq 2r_T. \tag{3.9}$$

The implications of Theorem 6 are as follows.

1. When the parameters η, ϵ, and $\sigma^{\mathrm{Q}}_{\max}$ do not scale with T, the rate of convergence reduces to $O_P(\sqrt{\log d/T})$. Thus, the number of assets under management is allowed to scale exponentially with sample size T. Compared to similar rates of convergence obtained for sample-covariance-based estimators (Bickel and Levina, 2008; Cai et al., 2010; Fan et al., 2013a), we do not require any moment or tail conditions, thus accommodating heavy-tailed asset return data.

2. The effect of serial dependence on the rate of convergence is characterized by C_ϵ. Specifically, as ϵ approaches 0, $C_\epsilon = \sum_{k=1}^{\infty} 1/k^{1+\epsilon}$ increases towards infinity, inflating r_T. ϵ is allowed to scale with T such that $C_\epsilon = o(T/\log d)$.

3. The rate of convergence r_T is inversely related to the lower bound, η, on the marginal

density functions around the $1/4$ quantiles. This is because when η is small, the distribution functions are flat around the $1/4$ quantiles, making the population quantiles harder to estimate.

Combining Lemma 3 and Theorem 6, we obtain the rate of convergence for $R(\tilde{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}})$.

Theorem 7. *Let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ be an absolutely continuous stationary process satisfying Conditions 1 and 2. Suppose that $\log d/T \to 0$ as $T \to \infty$ and $\mathbf{R}^{\mathrm{Q}} \in S_\lambda$. Then, for any $\alpha \in (0,1)$ and T large enough, we have*

$$|R(\tilde{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) - R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}})| \leq 2c^2 r_T, \tag{3.10}$$

where r_T is defined in (3.8) *and c is the gross-exposure constant.*

Theorem 7 shows that the risk of the estimated portfolio converges to the oracle optimal risk with parametric rate r_T. The number of assets, d, is allowed to scale exponentially with sample size T. Moreover, the rate of convergence does not rely on any tail conditions on the distribution of the asset returns.

For the rest of this section, we build the connection between the proposed robust portfolio optimization and its moment-based counterpart. Specifically, we show that they are consistent under the elliptical model.

Definition 8. (Fang et al., 1990) *A random vector $\boldsymbol{X} \in \mathbb{R}^d$ follows an elliptical distribution with location $\boldsymbol{\mu} \in \mathbb{R}^d$ and scatter $\mathbf{S} \in \mathbb{R}^{d\times d}$ if and only if there exist a nonnegative random variable $\xi \in \mathbb{R}$, a matrix $\mathbf{A} \in \mathbb{R}^{d\times r}$ with $\mathrm{rank}(\mathbf{A}) = r$, a random vector $\boldsymbol{U} \in \mathbb{R}^r$*

independent from ξ and uniformly distributed on the r-dimensional sphere, $\mathbb{S}^{r-1}$, such that

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \xi \mathbf{A}\boldsymbol{U}.$$

Here $\mathbf{S} = \mathbf{A}\mathbf{A}^{\mathsf{T}}$ has rank r. We denote $\boldsymbol{X} \sim \mathrm{EC}_d(\boldsymbol{\mu}, \mathbf{S}, \xi)$. ξ is called the generating variate.

Commonly used elliptical distributions include Gaussian distribution and t-distribution. Elliptical distributions have been widely used for modeling financial return data, since they naturally capture many stylized properties including heavy tails and tail dependence (Joe, 1997; Schmidt, 2002; Rachev, 2003; Rachev et al., 2005; Dowd, 2007; Andersen, 2009). The next theorem relates $\mathbf{R}^{\mathrm{Q}}$ and $R(\mathbf{w}; \mathbf{R}^{\mathrm{Q}})$ to their moment-based counterparts, $\boldsymbol{\Sigma}$ and $R(\mathbf{w}; \boldsymbol{\Sigma})$, under the elliptical model.

Theorem 9. *Let $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathsf{T}} \sim \mathrm{EC}_d(\mu, \mathbf{S}, \xi)$ be an absolutely continuous elliptical random vector and $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_d)^{\mathsf{T}}$ be an independent copy of $\boldsymbol{X}$. Then, we have*

$$\mathbf{R}^{\mathrm{Q}} = m^{\mathrm{Q}}\mathbf{S} \tag{3.11}$$

for some constant m^{Q} only depending on the distribution of $\boldsymbol{X}$. Moreover, if $0 < \mathbb{E}\xi^2 < \infty$,

we have

$$\mathbf{R}^{\mathrm{Q}} = c^{\mathrm{Q}}\mathbf{\Sigma} \;\; \textit{and} \;\; R(\mathbf{w}; \mathbf{R}^{\mathrm{Q}}) = c^{\mathrm{Q}} R(\mathbf{w}; \mathbf{\Sigma}), \tag{3.12}$$

where $\mathbf{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$ *is the covariance matrix of* $\boldsymbol{X}$*, and* c^{Q} *is a constant given by*

$$\begin{aligned} c^{\mathrm{Q}} =& Q\Big\{\frac{(X_j - \tilde{X}_j)^2}{\mathrm{Var}(X_j)}; \frac{1}{4}\Big\} = Q\Big\{\frac{(X_j + X_k - \tilde{X}_j - \tilde{X}_k)^2}{\mathrm{Var}(X_j + X_k)}; \frac{1}{4}\Big\} \\ =& Q\Big\{\frac{(X_j - X_k - \tilde{X}_j + \tilde{X}_k)^2}{\mathrm{Var}(X_j - X_k)}; \frac{1}{4}\Big\}. \end{aligned} \tag{3.13}$$

Here the last two inequalities hold when $\mathrm{Var}(X_j + X_k) > 0$ *and* $\mathrm{Var}(X_j - X_k) > 0$*.*

By Theorem 9, under the elliptical model, minimizing the robust risk metric, $R(\mathbf{w}; \mathbf{R}^{\mathrm{Q}})$, is equivalent with minimizing the standard moment-based risk metric, $R(\mathbf{w}; \mathbf{\Sigma})$. Thus, the robust portfolio optimization (3.3) is equivalent to its moment-based counterpart (3.1) in the population level. Plugging (3.12) into (3.10) leads to the following theorem.

Theorem 10. *Let* $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ *be an absolutely continuous stationary process satisfying Conditions 1 and 2. Suppose that* $\boldsymbol{X}_1 \sim \mathrm{EC}_d(\boldsymbol{\mu}, \mathbf{S}, \xi)$ *follows an elliptical distribution with covariance matrix* $\mathbf{\Sigma}$*, and* $\log d/T \to 0$ *as* $T \to \infty$*. Then, we have*

$$|R(\tilde{\mathbf{w}}^{\mathrm{opt}}; \mathbf{\Sigma}) - R(\mathbf{w}^{\mathrm{opt}}; \mathbf{\Sigma})| \leq \frac{2c^2}{c^{\mathrm{Q}}} r_T,$$

where c *is the gross-exposure constant,* c^{Q} *is defined in* (3.13)*, and* r_T *is defined in* (3.8)*.*

Thus, under the elliptical model, the optimal portfolio, $\tilde{\mathbf{w}}^{\mathrm{opt}}$, obtained from the robust portfolio optimization also leads to parametric rate of convergence for the standard moment-based risk.

3.5 Experiments

In this section, we investigate the empirical performance of the proposed portfolio optimization approach. In Section 3.5.1, we demonstrate the robustness of the proposed approach using synthetic heavy-tailed data. In Section 3.5.2, we simulate portfolio management using the Standard & Poor's 500 (S&P 500) stock index data.

The proposed portfolio optimization approach (QNE) is compared with three competitors. These competitors are constructed by replacing the covariance matrix $\mathbf{\Sigma}$ in (3.1) by commonly used covariance/scatter matrix estimators:

1. OGK: The orthogonalized Gnanadesikan-Kettenring estimator constructs a pilot scatter matrix estimate using a robust τ-estimator of scale, then re-estimates the eigenvalues using the variances of the principal components (Maronna and Zamar, 2002).

2. Factor: The principal factor estimator iteratively solves for the specific variances and the factor loadings (Bai and Shi, 2011).

3. Shrink: The shrinkage estimator shrinkages the sample covariance matrix towards a one-factor covariance estimator(Ledoit and Wolf, 2003).

3.5.1 Synthetic Data

Following Fan et al. (2012a), we construct the covariance matrix of the asset returns using a three-factor model:

$$X_j = b_{j1}f_1 + b_{j2}f_2 + b_{j3}f_3 + \varepsilon_j,\ j = 1, \ldots, d, \tag{3.14}$$

where X_j is the return of the j-th stock, b_{jk} is the loadings of the j-th stock on factor f_k, and ε_j is the idiosyncratic noise independent of the three factors. Under this model, the covariance matrix of the stock returns is given by

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_f\mathbf{B}^\mathsf{T} + \text{diag}(\sigma_1^2, \ldots, \sigma_d^2), \tag{3.15}$$

where $\mathbf{B} = [b_{jk}]$ is a $d \times 3$ matrix consisting of the factor loadings, $\mathbf{\Sigma}_f$ is the covariance matrix of the three factors, and σ_j^2 is the variance of the noise ε_i. We adopt the covariance in (3.15) in our simulations. Following Fan et al. (2012a), we generate the factor loadings $\mathbf{B}$ from a trivariate normal distribution, $N_d(\boldsymbol{\mu}_b, \mathbf{\Sigma}_b)$, where the mean, $\boldsymbol{\mu}_b$, and covariance, $\mathbf{\Sigma}_b$, are specified in Table 3.1. After the factor loadings are generated, they are fixed as parameters throughout the simulations. The covariance matrix, $\mathbf{\Sigma}_f$, of the three factors is also given in Table 3.1. The standard deviations, $\sigma_1, \ldots, \sigma_d$, of the idiosyncratic noises are generated independently from a truncated gamma distribution with shape 3.3586 and scale 0.1876, restricting the support to $[0.195, \infty)$. Again these standard deviations are fixed as

Table 3.1: Parameters for generating the covariance matrix in Equation (3.15).

Parameters for factor loadings				Parameters for factor returns		
$\boldsymbol{\mu}_b$	$\boldsymbol{\Sigma}_b$			$\boldsymbol{\Sigma}_f$		
0.7828	0.02915	0.02387	0.01018	1.2507	-0.035	-0.2042
0.5180	0.02387	0.05395	-0.00697	-0.0350	0.3156	-0.0023
0.4100	0.01018	-0.00697	0.08686	-0.2042	-0.0023	0.1930

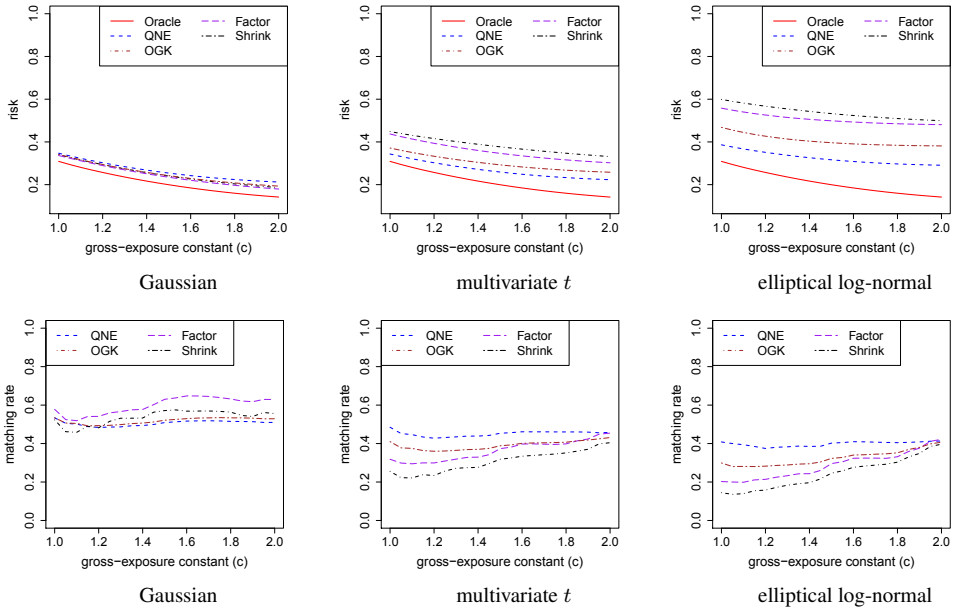


Figure 3.1: Portfolio risks, selected number of stocks, and matching rates to the oracle optimal portfolios.

parameters once they are generated. According to Fan et al. (2012a), these parameters are obtained by fitting the three-factor model, (3.14), using three-year daily return data of 30 Industry Portfolios from May 1, 2002 to Aug. 29, 2005. The covariance matrix, $\boldsymbol{\Sigma}$, is fixed throughout the simulations. Since we are only interested in risk optimization, we set the mean of the asset returns to be $\boldsymbol{\mu} = \mathbf{0}$. The dimension of the stocks under consideration is fixed at $d = 100$.

Given the covariance matrix $\boldsymbol{\Sigma}$, we generate the asset return data from the following

three distributions.

D_1: multivariate Gaussian distribution, $N_d(\mathbf{0}, \mathbf{\Sigma})$;

D_2: multivariate t distribution with degree of freedom 3 and covariance matrix $\mathbf{\Sigma}$;

D_2: elliptical distribution with log-normal generating variate, $\log N(0, 2)$, and covariance matrix $\mathbf{\Sigma}$.

Under each distribution, we generate asset return series of half a year ($T = 126$). We estimate the covariance/scatter matrices using QNE and the three competitors, and plug them into (3.1) to optimize the portfolio allocations. We also solve (3.1) with the true covariance matrix, $\mathbf{\Sigma}$, to obtain the oracle optimal portfolios as benchmarks. We range the gross-exposure constraint, c, from 1 to 2. The results are based on 1,000 simulations.

Figure 3.1 shows the portfolio risks $R(\hat{\mathbf{w}}; \mathbf{\Sigma})$ and the matching rates between the optimized portfolios and the oracle optimal portfolios[2]. Here the matching rate is defined as follows. For two portfolios P_1 and P_2, let S_1 and S_2 be the corresponding sets of selected assets, i.e., the assets for which the weights, w_i, are non-zero. The matching rate between P_1 and P_2 is defined as $r(P_1, P_2) = |S_1 \bigcap S_2|/|S_1 \bigcup S_2|$, where $|S|$ denotes the cardinality of set S.

We note two observations from Figure 3.1. (i) The four estimators leads to comparable portfolio risks under the Gaussian model D_1. However, under heavy-tailed distributions D_2 and D_3, QNE achieves lower portfolio risk. (ii) The matching rates of QNE are stable across the three models, and are higher than the competing methods under heavy-tailed

[2]Due to the ℓ_1 regularization in the gross-exposure constraint, the solution is generally sparse.

Table 3.2: Annualized Sharpe ratios, returns, and risks under 4 competing approaches, using S&P 500 index data.

		QNE	OGK	Factor	Shrink
Sharpe ratio	c=1.0	**2.04**	1.64	1.29	0.92
	c=1.2	**1.89**	1.39	1.22	0.74
	c=1.4	**1.61**	1.24	1.34	0.72
	c=1.6	**1.56**	1.31	1.38	0.75
	c=1.8	**1.55**	1.48	1.41	0.78
	c=2.0	**1.53**	1.51	1.43	0.83
return (in %)	c=1.0	**20.46**	16.59	13.18	9.84
	c=1.2	**18.41**	13.15	10.79	7.20
	c=1.4	**15.58**	11.30	10.88	6.55
	c=1.6	**15.02**	11.48	10.68	6.49
	c=1.8	**14.77**	12.39	10.57	6.58
	c=2.0	**14.51**	12.27	10.60	6.76
risk (in %)	c=1.0	**10.02**	10.09	10.19	10.70
	c=1.2	9.74	9.46	**8.83**	9.76
	c=1.4	9.70	9.10	**8.12**	9.14
	c=1.6	9.63	8.75	**7.71**	8.68
	c=1.8	9.54	8.39	**7.51**	8.38
	c=2.0	9.48	8.13	**7.43**	8.18

distributions D_2 and D_3. Thus, we conclude that QNE is robust to heavy tails in both risk minimization and asset selection.

3.5.2 Real Data

In this section, we simulate portfolio management using the S&P 500 stocks. We collect 1,258 adjusted daily closing prices[3] for 435 stocks that stayed in the S&P 500 index from January 1, 2003 to December 31, 2007. Using the closing prices, we obtain 1,257 daily returns as the daily growth rates of the prices.

We manage a portfolio consisting of the 435 stocks from January 1, 2003 to December 31, 2007[4]. On days $i = 42, 43, \ldots, 1,256$, we optimize the portfolio allocations using the

[3]The adjusted closing prices accounts for all corporate actions including stock splits, dividends, and rights offerings.

[4]We drop the data after 2007 to avoid the financial crisis, when the stock prices are likely to violate the

past 2 months stock return data (42 sample points). We hold the portfolio for one day, and evaluate the portfolio return on day $i + 1$. In this way, we obtain 1,215 portfolio returns. We repeat the process for each of the four methods under comparison, and range the gross-exposure constant c from 1 to 2[5].

Since the true covariance matrix of the stock returns is unknown, we adopt the Sharpe ratio for evaluating the performances of the portfolios. Table 3.2 summarizes the annualized Sharpe ratios, mean returns, and empirical risks (i.e., standard deviations of the portfolio returns). We observe that QNE achieves the largest Sharpe ratios under all values of the gross-exposure constant, indicating the lowest risks under the same returns (or equivalently, the highest returns under the same risk).

3.6 Discussion

In this paper, we propose a robust portfolio optimization framework, building on a quantile-based scatter matrix. We obtain non-asymptotic rates of convergence for the scatter matrix estimators and the risk of the estimated portfolio. The relations of the proposed framework with its moment-based counterpart are well understood.

The main contribution of the robust portfolio optimization approach lies in its robustness to heavy tails in high dimensions. Heavy tails present unique challenges in high dimensions compared to low dimensions. For example, asymptotic theory of M-estimators

stationary assumption.

[5] $c = 2$ imposes a 50% upper bound on the percentage of short positions. In practice, the percentage of short positions is usually strictly controlled to be much lower.

guarantees consistency in the rate $O_P(\sqrt{d/n})$ even for non-Gaussian data (Van De Geer and Van De Geer, 2000; Hall, 2005). If $d \ll n$, statistical error diminishes rapidly with increasing n. However, when $d \gg n$, statistical error may scale rapidly with dimension. Thus, stringent tail conditions, such as subGaussian conditions, are required to guarantee consistency for moment-based estimators in high dimensions (Bühlmann and Van De Geer, 2011). In this paper, based on quantile statistics, we achieve consistency for portfolio risk without assuming any tail conditions, while allowing d to scale nearly exponentially with n.

Another contribution of his work lies in the theoretical analysis of how serial dependence may affect consistency of the estimation. We measure the degree of serial dependence using the ϕ-mixing coefficient, $\phi(n)$. We show that the effect of the serial dependence on the rate of convergence is summarized by the parameter C_ϵ, which characterizes the size of $\sum_{n=1}^{\infty} \phi(n)$.

Chapter 4

A Theory of Kolmogorov Dependence with Applications to Scatter Matrix Estimation

4.1 Introduction

Dependent data arises from a wide range of applications. For example, in finance, the series of asset returns commonly exhibit short-term or long-term memory (Andersen, 2009); in functional magnetic resonance imaging (fMRI), the images from neighboring scans are serially dependent (Purdon and Weisskoff, 1998; Woolrich et al., 2001); in geophysics, data measured in geographical sites usually exhibit temporal dependence (Majda and Wang, 2006).

The prevalence of serial dependence has motivated the development of various dependence assumptions. These assumptions can be categorized into structural assumptions and non-structural assumptions. The former are based on specific models for the data generating mechanism. Examples of structural assumptions include vector autoregressive (VAR) models and physical dependence conditions. A brief review of these conditions and their applications is as follows.

- **VAR models**: The VAR models specify that the observed random vector depends linearly on its previous realizations. Under this model, Loh and Wainwright (2012) considered sparse linear regression; Han and Liu (2013b) proposed to estimate the transition matrix via a Dantzig-selector-type approach; Wang et al. (2013) studied the performance of sparse principal component analysis; Qiu et al. (2015) considered estimating time varying graphical models.

- **Physical dependence**: For stationary causal processes in the form of $\{\boldsymbol{X}_t = g(\{\boldsymbol{\epsilon}_j\}_{j\leq t})$

$\}_{t\in\mathbb{Z}}$, the physical dependence condition (Wu, 2005) assumes that the dependence strength between $\boldsymbol{X}_t = g(\{\boldsymbol{\epsilon}_j\}_{j\leq t})$ and $\boldsymbol{X}'_t = g(\{\boldsymbol{\epsilon}'_0, \boldsymbol{\epsilon}_j : j \leq t, j \neq 0\})$ decays to 0 as t goes to infinity. Here $\{\boldsymbol{\epsilon}'_0, \boldsymbol{\epsilon}_j : j \in \mathbb{Z}\}$ is a sequence of independent and identically distributed random vectors, and g is a measurable function[1]. Under this condition, Xiao and Wu (2012) derived rates of convergence for banding and thresholding estimators of the autocovariance matrix for stationary time series. Chen et al. (2013) considered estimation of covariance and inverse covariance matrices for stationary and locally stationary time series.

Despite the wide applications of the structural dependence assumptions, their main inconvenience is that they are often difficult to verify for a general process where the generating mechanism is unknown[2]. In contrast, non-structural dependence conditions rely on model-free dependence measures. For a time series $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$, these dependence measures quantify the degree of dependence between the "past", $\{\boldsymbol{X}_t\}_{t\leq 0}$, and the "future", $\{\boldsymbol{X}_t\}_{t\geq n}$. Examples of non-structural dependence conditions include the mixing conditions and the weak dependence conditions. A brief review on these conditions and the related applications is as follows.

- **Mixing conditions**: The mixing conditions are built on various mixing coefficients, which quantify the dependence strength between the σ-fields generated by $\{\boldsymbol{X}_t\}_{t\leq 0}$

[1] $\boldsymbol{X}_t = g(\{\boldsymbol{\epsilon}_j\}_{j\leq t})$ is interpreted as a physical system with $\{\boldsymbol{\epsilon}_j\}_{j\leq t}$ as the inputs and $\boldsymbol{X}_t$ as the output.

[2] We note that the data generating mechanisms themselves can be fairly general. For example, linear processes are special cases of stationary causal processes with $g(\{\boldsymbol{\epsilon}_j\}_{j\leq t}) = \sum_{k=0}^{\infty} \boldsymbol{\Phi}_k \boldsymbol{\epsilon}_{t-k}$, where $\boldsymbol{\Phi}_0 = \mathbf{I}_d$ and $\boldsymbol{\Phi}_k \in \mathbb{R}^{d\times d}$. Wold's decomposition theorem (Wold, 1938) states that any process where the only deterministic term is the mean term can be represented as a linear process.

and $\{\boldsymbol{X}_t\}_{t\geq n}$. The mixing conditions specify that the mixing coefficients decay to 0 as n goes to infinity. Assuming exponentially decaying α-mixing coefficients, Fan et al. (2012b) studied the asymptotic behavior of the sample covariance matrix. Fan et al. (2011) and Fan et al. (2013a) considered covariance matrix estimation under factor models with factors observed and unobserved, respectively. Based on these covariance matrix estimators, Fan et al. (2013b) derived limiting distributions for portfolio risk estimators. Bai and Liao (2012) and Bai and Liao (2013) derived limiting distributions for the estimated factors and factor loadings. Besides the α-mixing conditions, Pan and Yao (2008) and Lam et al. (2011) exploited the ϕ- and ψ-mixing conditions in estimating factors and factor loadings. Han and Liu (2013a) studied principal component analysis under the ϕ- and η-mixing conditions.

- **Weak dependence**: The weak dependence conditions rely on a dependence measure quantified by the covariance between smooth functions of $\{\boldsymbol{X}_t\}_{t\leq 0}$ and $\{\boldsymbol{X}_t\}_{t\geq n}$, and require that the covariance goes to 0 as n goes to infinity (Doukhan and Louhichi, 1999). Under the weak dependence conditions, Kallabis and Neumann (2006) and Doukhan and Neumann (2007) derived various probability and moment inequalities for weakly dependent processes; Fan et al. (2012b) studied the sample covariance matrix; Sancetta (2008) considered shrinkage estimators of covariance matrices.

The mixing conditions have been criticized for being difficult to verify (Doukhan and Louhichi, 1999). The difficulty is mainly due to the complex σ-fields involved in the definitions of the mixing coefficients. In comparison, the weak dependence conditions are

easier to verify in many scenarios. However, the covariance-based dependence measure only considers smooth transformations of the data. These conditions are not directly applicable to many other scenarios, such as the analysis of many quantile-based statistics, where non-smooth transformations are involved.

In this paper, we develop a new dependence measure named the Kolmogorov dependence measure. The dependence measure is naturally formulated using the Kolmogorov distance. Specifically, for two sequences of random variables, we quantify their dependence by the Kolmogorov distance between their joint distribution and the product of their marginal distributions. Using this dependence measure, we develop the Kolmogorov dependence condition for multivariate time series. We reveal its connections with VAR models, mixing conditions, physical dependence, and several covariance-based dependence conditions, and show that it's weaker than many commonly used dependence conditions.

The main challenge in building the connections between the Kolmogorov dependence condition and other conditions lies in the fundamental difference in the dependence measures. In particular, the Kolmogorov dependence measure is essentially the covariance between non-smooth transformations of the data. Standard techniques for analyzing smooth transformations no longer apply. To overcome the difficulty, we develop a set of techniques based on a novel construction of smooth functions for approximating given discontinuous ones. These techniques enables the verification of the Kolmogorov dependence condition under a wide variety of existing dependence conditions.

To demonstrate the importance of the Kolmogorov dependence condition, we analyze

a family of quantile-based scatter matrix estimators under dependent data. In particular, we demonstrate that the Kolmogorov dependence measure is naturally coupled with the structure of these estimators. This enables us to obtain fast rates of convergence for these estimators under the Kolmogorov dependence condition. The rates of convergence characterizes the impact of serial dependence on the consistency of the estimators. Since the Kolmogorov dependence condition is weaker than a number of other dependence conditions, rates of convergence of the scatter matrix estimators can be immediately obtained under these other dependence conditions as well.

Our contributions are three-fold. First, we propose a novel dependence condition with a novel dependence measure. Its connections with widely used dependence conditions are well understood. Secondly, under the Kolmogorov dependence condition, we derive optimal rates of convergence for a family of quantile-based scatter matrix estimators. Prior to this work, the performance of these estimators under dependent data is unknown. Lastly, we develop a set of techniques for analyzing time series characterized by the Kolmogorov dependence condition. These techniques are of independent interest in analyzing weakly dependent time series.

4.1.1 Organization

We organize the rest of this paper as follows. In Section 4.2, we propose the Kolmogorov dependence condition, and discuss its relations with other weak dependence conditions. In Section 4.3, we apply the Kolmogorov dependence condition to analyzing a

family of quantile-based scatter matrix estimators. We gather the proofs of the main theoretical results in Section 4.4. In Section 4.5, we summarize the main contributions of this paper. Additional technical results are collected in the Appendix C.

4.1.2 Notation

Let $\mathbf{v} = (v_1, \ldots, v_d)^\mathsf{T}$ be a d-dimensional real vector, and $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{d_1 \times d_2}$ be a $d_1 \times d_2$ matrix with M_{jk} as the (j,k) entry. For $0 < q < \infty$, we define the ℓ_q vector norm of $\mathbf{v}$ as $\|\mathbf{v}\|_q := (\sum_{j=1}^d |v_j|)^{1/q}$ and the ℓ_∞ vector norm of $\mathbf{v}$ as $\|\mathbf{v}\|_\infty := \max_j |v_j|$. Let the matrix $\ell_{\max}$ norm of $\mathbf{M}$ to be $\|\mathbf{M}\|_{\max} := \max_{jk} |M_{jk}|$, the matrix ℓ_∞ norm of $\mathbf{M}$ be $\|\mathbf{M}\|_\infty = \max_j \sum_{k=1}^d |M_{jk}|$, and the Frobenius norm to be $\|\mathbf{M}\|_F := \sqrt{\sum_{jk} M_{jk}^2}$. We define $\mathrm{vec}(\mathbf{M})$ to be the vector obtained by stacking the columns of $\mathbf{M}$:

$$\mathrm{vec}(\mathbf{M}) := (M_{11}, \ldots, M_{d_1 1}, M_{12}, \ldots, M_{d_1 2}, \ldots, M_{1 d_2}, \ldots, M_{d_1 d_2})^\mathsf{T}.$$

Conversely, define $\mathrm{mat}\{\mathrm{vec}(\mathbf{M})\} := \mathbf{M}$ as the original matrix $\mathbf{M}$. Let $\mathbf{N} = [N_{jk}]$ be another matrix with the same dimension as $\mathbf{M}$. We denote the Hadamard product of $\mathbf{M}$ and $\mathbf{N}$ as $\mathbf{M} \circ \mathbf{N} := [M_{jk} N_{jk}]$. We denote $\mathbf{M} \preceq \mathbf{N}$ if $\mathbf{N} - \mathbf{M}$ is positive semi-definite.

For a sequence of numbers $a_1, \ldots, a_d$, we denote $\mathrm{diag}(a_1, \ldots, a_d)$ to be a diagonal matrix with diagonal entries $a_1, \ldots, a_d$. Similarly, for a sequence of matrices $\mathbf{A}_1, \ldots, \mathbf{A}_d$, we denote $\mathrm{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_d)$ to be a block diagonal matrix with diagonal blocks $\mathbf{A}_1, \ldots, \mathbf{A}_d$. Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\mathsf{T}$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\mathsf{T}$ be two random vectors. We write $\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{Y}$

if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed. Let $\mathcal{S}, \mathcal{T} \subseteq \{1, \ldots, T\}$ be two index sets. We denote $|\mathcal{S}|$ as the cardinality of $\mathcal{S}$, and $d(\mathcal{S}, \mathcal{T}) := \inf\{|s - t| : s \in \mathcal{S}, t \in \mathcal{T}\}$ as the minimal distance between the elements in $\mathcal{S}$ and $\mathcal{T}$. For $a, b \in \mathbb{R}$, let $a \vee b := \max\{a, b\}$. Throughout the paper, we use $C, C_1, C_2, \ldots$ to denote generic constants, though the actual values may vary at different occasions. We use $\mathbf{1}, \mathbf{2}, \ldots$ to denote vectors with $1, 2, \ldots$ at every entry.

4.2 Kolmogorov Dependence

We first introduce a measure of dependence between two sequences based on the Kolmogorov distance.

Definition 11. *Let $\{X_s\}_{s\in\mathcal{S}}$ and $\{Y_t\}_{t\in\mathcal{T}}$ be two sequences of random variables indexed by sets $\mathcal{S}, \mathcal{T} \subseteq \mathbb{Z}$. We define the Kolmogorov dependence measure between the two sequences by*

$$\boldsymbol{\kappa}(\{X_s\}_{s\in\mathcal{S}}, \{Y_t\}_{t\in\mathcal{T}}) :=$$
$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big(X_s \le u, Y_t \le u, \forall s \in \mathcal{S}, t \in \mathcal{T}\Big) - \mathbb{P}\Big(X_s \le u, \forall s \in \mathcal{S}\Big)\mathbb{P}\Big(Y_t \le u, \forall t \in \mathcal{T}\Big)\Big|.$$

If we define $F(u) := \mathbb{P}(X_s \le u, Y_t \le u, \forall s \in \mathcal{S}, t \in \mathcal{T})$ and $G(u) := \mathbb{P}(X_s \le u, \forall s \in \mathcal{S})\mathbb{P}(Y_t \le u, \forall t \in \mathcal{T})$, the Kolmogorov dependence measure between $\{X_s\}_{s\in\mathcal{S}}$ and $\{Y_t\}_{t\in\mathcal{T}}$ is the Kolmogorov distance between F and G: $\boldsymbol{\kappa}(\{X_s\}_{s\in\mathcal{S}}, \{Y_t\}_{t\in\mathcal{T}}) = \sup_{u\in\mathbb{R}}|F(u)$

$- G(u)|$. Based on the Kolmogorov dependence measure, we next introduce the Kolmogorov dependence condition for modeling the serial dependence in multivariate time series.

Condition 3. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ be a stationary sequence of random vectors. Let $\Psi : \mathbb{N}^2 \to \mathbb{N}$ be any one*[3] *of the following four functions:*

(a) $\Psi(u, v) = 2v$,

(b) $\Psi(u, v) = u + v$,

(c) $\Psi(u, v) = uv$,

(d) $\Psi(u, v) = \beta(u + v) + (1 - \beta)uv$, *for some* $\beta \in (0, 1)$.

The sequence $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ satisfies the Kolmogorov dependence condition if and only if the following two requirements are satisfied:

1. *There exist a constant $K > 0$ and a real sequence $\{\rho(n)\}_{n\geq 0}$ such that for any non-empty sets $\mathcal{S}, \mathcal{T} \subseteq \{1, \ldots, T\}$ with $\max(\mathcal{S}) \leq \min(\mathcal{T})$, and any sequence $\{Y_t\}_{t=1}^T \in \{\{X_{tj}\}_{t=1}^T, \{X_{tj} + X_{tk}\}_{t=1}^T, \{X_{tj} - X_{tk}\}_{t=1}^T : j \neq k \in \{1, \ldots, d\}\}$, we have*

$$\boldsymbol{\kappa}(\{Y_s\}_{s\in\mathcal{S}}, \{Y_t\}_{t\in\mathcal{T}}) \leq K^2\Psi(|\mathcal{S}|, |\mathcal{T}|)\rho\{d(\mathcal{S}, \mathcal{T})\}.$$

[3] We only require that Condition 3 holds for at least one of the four Ψ functions.

2. *The sequence* $\{\rho(n)\}_{n\geq 0}$ *satisfies*

$$\sum_{n=0}^{\infty}(n+1)^k\rho(n) \leq L_1 L^k (k!)^a, \text{ for any } k \geq 0 \text{ and } k \in \mathbb{Z}, \tag{4.1}$$

where $L_1 > 0$ *and* $a \geq 0$ *are constants and* L *may scale with* (T, d) *such that*

$$L = L(T,d) \leq \frac{K\sqrt{L_1}}{2^{7a/2+6}\sqrt{K^2 \vee 2}} \cdot \frac{\sqrt{T}}{(\log d)^{a+3/2}}. \tag{4.2}$$

The sequence $\{\rho(n)\}_{n\geq 0}$ characterizes the decay of dependence strength, measured by $\boldsymbol{\kappa}$, over time. Equation (4.1) specifies the desired rate of decay. The upper bound in (4.1) is adaptive to the sample size T and dimension d, in the sense that L is allowed to scale with (T, d) by the rate $\sqrt{T}/(\log d)^{a+3/2}$. Intuitively, larger sample size provides more information, which in turn allows for stronger dependence among the sample. On the other hand, larger dimension of the data entails weaker dependence. Overall, d is allowed to scale in the rate $\exp\{T^{1/(2a+3)}\}$ without collapsing L to 0.

In the following, we unveil the relation between the Kolmogorov dependence condition and several weak dependence conditions frequently exploited in the literature. In particular, we show that many time series satisfying certain dependence conditions (VAR models, α-mixing conditions, weak dependence, and physical dependence) also satisfy Condition 3.

Theorem 12 (VAR model)**.** *Let* $\{\boldsymbol{X}_t \in \mathbb{R}^d\}_{t\in\mathbb{Z}}$ *be a stationary process satisfying the vector*

autoregressive model

$$\boldsymbol{X}_t = \mathbf{A}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \text{ for any } t \in \mathbb{Z},$$

where $\{\boldsymbol{\epsilon}_t\}_{t\in\mathbb{Z}}$ *is a sequence of i.i.d. random vectors. Assume the following conditions hold:*

1. $\|\mathbf{A}\|_2 < 1$.
2. $\mathbb{E}|\mathbf{e}_j^\mathsf{T}\mathbf{A}^\ell\boldsymbol{\epsilon}_1| \le C\|\mathbf{A}\|_2^\ell$ *for* $j = 1, \ldots, d$, *any* $\ell \in \mathbb{Z}^+$, *and some positive constant* C, *where* $\mathbf{e}_j$ *is the* j*-th column of the identity matrix.*
3. *There exists a constant* $H > 0$ *such that* $\mathbb{P}(u \le Y \le u + v) \le Hv$ *for any* $u \in \mathbb{R}$, $v > 0$, *and* $Y \in \{X_{1j}, X_{1j} + X_{1k}, X_{1j} - X_{1k} : j, k = 1, \ldots, d\}$.

Then Condition 3 holds for the sequence $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ *with* $a = 1$, $\Psi(u, v) = u + v$, $K = 4H + 3C/(1 - \|\mathbf{A}\|_2)$, *and* $L_1 = L = 1/(1 - \sqrt{\|\mathbf{A}\|_2})$.

Remark 13. The first assumption guarantees that $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is a stable process. The third assumption is a smoothness condition on the marginal distribution functions. For the second assumption, when d is fixed, since $\mathbb{E}|\mathbf{e}_j^\mathsf{T}\mathbf{A}^\ell\boldsymbol{\epsilon}_1| \le \|\mathbf{e}_j^\mathsf{T}\mathbf{A}^\ell\|_2\mathbb{E}\|\boldsymbol{\epsilon}_1\|_2 \le \|\mathbf{A}\|_2^\ell\mathbb{E}\|\boldsymbol{\epsilon}_1\|_2$, the assumption is satisfied provided that $\mathbb{E}\|\boldsymbol{\epsilon}_1\|_2 < \infty$. When d may scale with sample size T, this assumption can be satisfied by assuming either Gaussian innovations, $\{\boldsymbol{\epsilon}_t\}_{t\in\mathbb{Z}}$, or certain sparsity structures on the transition matrix $\mathbf{A}$:

1. Gaussian innovations: Suppose that $\boldsymbol{\epsilon}_1 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ follows a Gaussian distribution with $\|\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\|_2 \leq C$ for some constant C. By the properties of Gaussian distributions, we have

$$\mathbf{e}_j^\mathsf{T} \mathbf{A}^\ell \boldsymbol{\epsilon}_1 \sim N(0, \mathbf{e}_j^\mathsf{T} \mathbf{A}^\ell \boldsymbol{\Sigma}_\epsilon \mathbf{A}^{\ell^\mathsf{T}} \mathbf{e}_j).$$

Thus, we have

$$\mathbb{E}|\mathbf{e}_j^\mathsf{T} \mathbf{A}^\ell \boldsymbol{\epsilon}_1| = \sqrt{\frac{2}{\pi} \mathbf{e}_j^\mathsf{T} \mathbf{A}^\ell \boldsymbol{\Sigma}_\epsilon \mathbf{A}^{\ell^\mathsf{T}} \mathbf{e}_j} \leq \sqrt{\frac{2}{\pi} \|\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\|_2 \|\mathbf{A}^{\ell^\mathsf{T}} \mathbf{e}_j\|_2^2} \leq \sqrt{\frac{2}{\pi} C} \|\mathbf{A}\|_2^\ell.$$

2. Sparse transition matrix: Suppose that $\mathbf{A}$ is block diagonal: $\mathbf{A} = \mathrm{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_m)$, where $d_i := \dim(\mathbf{A}_i)$ is fixed for $i = 1, \ldots, m$ while m may scale with T. In other words, $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ consists of autoregressive blocks. In this case, let $i_0 = \min\{i : j \leq d_i\}$ and partition $\boldsymbol{\epsilon}_1 = (\boldsymbol{\epsilon}_{11}, \ldots, \boldsymbol{\epsilon}_{1m})$ according to the dimensions of $(\mathbf{A}_1, \ldots, \mathbf{A}_m)$. We have $\mathbb{E}|\mathbf{e}_j^\mathsf{T} \mathbf{A}^\ell \boldsymbol{\epsilon}_1| \leq \|\mathbf{A}_{i_0}\|_2^\ell \mathbb{E}\|\boldsymbol{\epsilon}_{1 i_0}\|_2 \leq \|\mathbf{A}\|_2^\ell \mathbb{E}\|\boldsymbol{\epsilon}_{1 i_0}\|_2$. Thus, the second assumption is satisfied if $\mathbb{E}\|\boldsymbol{\epsilon}_{1i}\|_2 < \infty$ for $i = 1, \ldots, m$.

Next, we introduce the α-mixing process.

Definition 14 (Bradley (2005))**.** *Let $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ be a stationary stochastic process. For $-\infty \leq J \leq L \leq \infty$, define $\mathscr{F}_J^L := \sigma(\boldsymbol{X}_t : J \leq t \leq L, t \in \mathbb{Z})$ as the σ-field generated by $\{\boldsymbol{X}_t : J \leq t \leq L, t \in \mathbb{Z}\}$. For any $n \geq 1$, we define the α-mixing coefficient*

as

$$\boldsymbol{\alpha}(n) := \sup_{A \in \mathscr{F}_{-\infty}^{0}, B \in \mathscr{F}_{n}^{\infty}} \Big| \mathbb{P}(A \cap B) - P(A)P(B) \Big|.$$

The process $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ *is* α*-mixing if and only if* $\lim_{n \to \infty} \boldsymbol{\alpha}(n) = 0$.

The mixing coefficient $\boldsymbol{\alpha}(n)$ measures the dependence of two subsequences with index gap n. The rate at which $\boldsymbol{\alpha}(n)$ converges to 0 characterizes the degree of dependence over the process. If $\boldsymbol{\alpha}(n) = 0$ for all n, the process $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ is independent.

Theorem 15 (α-mixing)**.** *Let* $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ *be an* α*-mixing process with exponentially decaying* α*-mixing coefficient:*

$$\boldsymbol{\alpha}(n) \leq C_1 \exp\Big(-C_2 n^r\Big), \tag{4.3}$$

where $C_1, C_2, r > 0$ *are constants. Then Condition 3 holds for the sequence* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ *with* $a = \max(1, 1/r)$*, constants* K*,* L_1*, and* L *only depending on* C_1*,* C_2*, and* r*, and any of the four* Ψ *functions.*

Theorem 15 shows that Condition 3 is weaker than the exponentially decaying α-mixing condition (4.3). Condition (4.3) has been heavily exploited in modeling dependence in financial time series. See, for example, Fan et al. (2011), Fan et al. (2012b), Fan et al. (2013a), Fan et al. (2013b), Bai and Liao (2012), and Bai and Liao (2013) among others.

Compared to the α-mixing condition, Condition 3 is easier to verify. For example,

investigating the relation between VAR models and the α-mixing condition has proven to be difficult, mainly due to the complicated σ-fields involved in the definition of the mixing coefficient (Chanda, 1974; Gorodetskii, 1978; Andrews, 1984; Pham and Tran, 1985). In comparison, the proof of Theorem 12 is natural and concise.

Next, we introduce Doukhan's weak dependence measure (Doukhan and Louhichi, 1999). For a function $g : (\mathbb{R}^d)^u \to \mathbb{R}$, we define

$$\mathrm{Lip}(g) := \sup\Big\{ \frac{|g(\mathbf{x}_1, \ldots, \mathbf{x}_u) - g(\mathbf{y}_1, \ldots, \mathbf{y}_u)|}{\|\mathbf{x}_1 - \mathbf{y}_1\|_q + \cdots + \|\mathbf{x}_u - \mathbf{y}_u\|_q} : (\mathbf{x}_1, \ldots, \mathbf{x}_u) \neq (\mathbf{y}_1, \ldots, \mathbf{y}_u) \Big\},$$

where $0 < q \leq \infty$ is a constant. Denote $\Lambda := \{g : (\mathbb{R}^u)^d \to \mathbb{R} \text{ for some } u \ : \mathrm{Lip}(g) < \infty\}$ and $\Lambda^{(1)} := \{g \in \Lambda : \|g\|_\infty \leq 1\}$, where $\|g\|_\infty := \sup_{\mathbf{x}} g(\mathbf{x})$.

Definition 16 (Doukhan and Louhichi (1999); Doukhan and Neumann (2007))**.** *The process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is $(\Lambda^{(1)}, \psi, \zeta)$-weakly dependent if and only if there exists a function $\psi : \mathbb{R}_+^2 \times \mathbb{N}^2 \to \mathbb{R}_+$ and a sequence $\zeta = \{\zeta(n)\}_{n\geq 0}$ decreasing to 0 as n goes to infinity, such that for any $g_1, g_2 \in \Lambda^{(1)}$ with $g_1 : (\mathbb{R}^d)^u \to \mathbb{R}$, $g_2 : (\mathbb{R}^d)^v \to \mathbb{R}$, $u, v \in \mathbb{N}$, and any u-tuple $(s_1, \ldots, s_u)$ and any v-tuple $(t_1, \ldots, t_v)$ with $s_1 \leq \cdots \leq s_u < t_1 \leq \cdots \leq t_v$, the following inequality is satisfied:*

$$\Big|\mathrm{Cov}\Big\{ g_1(\boldsymbol{X}_{s_1}, \ldots, \boldsymbol{X}_{s_u}), g_2(\boldsymbol{X}_{t_1}, \ldots, \boldsymbol{X}_{t_v}) \Big\}\Big| \leq \psi(\mathrm{Lip}(g_1), \mathrm{Lip}(g_2), u, v)\zeta(t_1 - s_u).$$

Important examples of $(\Lambda^{(1)}, \psi, \zeta)$-weakly dependent processes include θ-, η-, κ-, and λ-dependence, which are listed in Table 4.1. They correspond to specific choices of the func-

Table 4.1: Important examples of weak dependence.

θ-dependence:	$\psi(\text{Lip}g_1, \text{Lip}g_2, u, v) = v\text{Lip}(g_2)$
η-dependence:	$\psi(\text{Lip}g_1, \text{Lip}g_2, u, v) = u\text{Lip}(g_1) + v\text{Lip}(g_2)$
κ-dependence:	$\psi(\text{Lip}g_1, \text{Lip}g_2, u, v) = uv\text{Lip}(g_1)\text{Lip}(g_2)$
λ-dependence:	$\psi(\text{Lip}g_1, \text{Lip}g_2, u, v) = u\text{Lip}(g_1) + v\text{Lip}(g_2) + uv\text{Lip}(g_1)\text{Lip}(g_2)$

tion ψ.

Similar to the α-mixing coefficient, the sequence ζ describes the degree of dependence over the process. The next theorem relates the weak dependence to Condition 3.

Theorem 17 (Weak dependence)**.** *Let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ be a $(\Lambda^{(1)}, \psi, \zeta)$-weakly dependent stationary process. Suppose there exists a constant $H > 0$ such that $\mathbb{P}(u \leq Y \leq u + v) \leq Hv$ holds for any $u \in \mathbb{R}$, $v > 0$, and $Y \in \{X_{1j}, X_{1j} + X_{1k}, X_{1j} - X_{1k} : j, k = 1, \ldots, d\}$. Then the following statements hold:*

1. *If $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is θ- or η-dependent and the sequence $\{\rho(n) = \sqrt{\zeta(n)}\}_{n\geq 0}$ satisfies* (4.1)*, Condition 3 holds for the sequence $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ with $\Psi(u, v) = u + v$.*

2. *If $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is κ- or λ-dependent and the sequence $\{\rho(n) = \zeta(n)^{1/3}\}_{n\geq 0}$ satisfies* (4.1)*, Condition 3 holds for the sequence $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ with $\Psi(u, v) = \beta(u+v)+(1-\beta)uv$, where $\beta = 16H/(16H+9)$ for κ-dependence and $\beta = (16H+6)/(16H+15)$ for λ-dependence.*

Next, we introduce the notion of m-dependence.

Definition 18. *The process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is m-dependent if and only if for any $t \in \mathbb{Z}$, $\{\boldsymbol{X}_s : s \leq t\}$ and $\{\boldsymbol{X}_s : s > t + m\}$ are independent.*

If the process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is m-dependent, it's α-mixing with $\boldsymbol{\alpha}(n) = 0$ whenever $n > m$, and $(\Lambda^{(1)}, \psi, \zeta)$-weakly dependent with $\zeta(n) = 0$ whenever $n > m$. Thus, we have the following corollary.

Corollary 19 (m-dependence)**.** *Condition 3 is satisfied by any m-dependent process.*

Lastly, we introduce the physical dependence measure introduced in Wu (2005).

Definition 20 (Wu (2005))**.** *Let $\{\boldsymbol{\epsilon}_t\}_{t\in\mathbb{Z}}$ be i.i.d. random vectors, and $\{\boldsymbol{\epsilon}'_t\}_{t\in\mathbb{Z}}$ be an i.i.d. copy of $\{\boldsymbol{\epsilon}_t\}_{t\in\mathbb{Z}}$. For a set $I \subseteq \mathbb{Z}$, let $\boldsymbol{\epsilon}_{t,I} := \boldsymbol{\epsilon}'_t$ if $t \in I$ and $\boldsymbol{\epsilon}_{t,I} := \boldsymbol{\epsilon}_t$ if $t \notin I$. Let $\mathcal{F}_t := \{\ldots, \boldsymbol{\epsilon}_{t-1}, \boldsymbol{\epsilon}_t\}$ be a shift process, and $\mathcal{F}_{t,I} := \{\ldots, \boldsymbol{\epsilon}_{t-1,I}, \boldsymbol{\epsilon}_{t,I}\}$ be a coupled version of $\mathcal{F}_t$, where $\boldsymbol{\epsilon}_t$ is replaced by $\boldsymbol{\epsilon}'_t$ if $t \in I$. Let g be a measurable function. We define the physical dependence measure to be*

$$\delta(I, t, g) := \mathbb{E}|g(\mathcal{F}_t) - g(\mathcal{F}_{t,I})|.$$

The process $\{X_t = g(\mathcal{F}_t)\}_{t\in\mathbb{Z}}$ is stationary, and is causal or non-anticipative in the sense that X_t does not depend on future innovations $\{\boldsymbol{\epsilon}_s : s > t\}$. $\mathcal{F}_t$ and X_t can be regarded as the inputs and output of a physical system g. The next theorem gives sufficient conditions for a multivariate physical process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ to satisfy Condition 3.

Theorem 21. *Let $\boldsymbol{g} = (g_1, \ldots, g_d)^\mathsf{T}$ be an $\mathbb{R}^d$-valued measurable function and $\boldsymbol{X}_t = \boldsymbol{g}(\mathcal{F}_t) = (g_1(\mathcal{F}_t), \ldots, g_d(\mathcal{F}_t))^\mathsf{T}$. Let $I = \{0, -1, -2, \ldots\}$ and define $\theta_{t,j} := \delta(I, t, g_j)$. Assume the following conditions hold:*

1. *The sequence* $\rho(n) := \max_{j=1,\ldots,d} \sqrt{\theta_{n,j}}$ *satisfies* (4.1).

2. *There exists a constant* $H > 0$ *such that* $\mathbb{P}(u \leq Y \leq u + v) \leq Hv$ *for any* $u \in \mathbb{R}$, $v > 0$, *and* $Y \in \{X_{1j}, X_{1j} + X_{1k}, X_{1j} - X_{1k} : j, k = 1, \ldots, d\}$.

Then the sequence $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ *satisfies Condition 3 with* $\Psi(u, v) = u + v$.

4.3 Robust Scatter Matrix Estimation

In this section, we apply the Kolmogorov dependence condition to analyzing a family of robust scatter matrix estimators under time series data. We show that the Kolmogorov dependence condition is naturally coupled with the structure of these estimators, and enables us to characterize the effect of serial dependence on their rates of convergence.

Let $Z \in \mathbb{R}$ be a random variable and $q \in [0, 1]$ be a constant. We define the q-quantile of Z as

$$Q(Z; q) := \inf\{z : \mathbb{P}(Z \leq z) \geq q\}.$$

$Q(Z; q)$ is unique if there exists a unique z such that $\mathbb{P}(Z \leq z) = q$. Correspondingly, we define the empirical q-quantile of a sample, $\{z_t\}_{t=1}^T$, as

$$\hat{Q}(\{z_t\}; q) := z^{(k)}, \text{ where } k = \min\Big\{t : \frac{t}{T} \geq q\Big\}. \tag{4.4}$$

Here $z^{(1)} \leq z^{(2)} \leq \cdots \leq z^{(T)}$ are the order statistics of $z_1, \ldots, z_T$. Building on quantiles,

the median absolute deviation (MAD) (Hampel, 1974) provides a robust measure of scales. The population and sample MADs are defined as[4]

$$\sigma^{\mathrm{M}}(Z;q) := Q\Big(\Big\{\Big|Z - Q\Big(Z;\frac{1}{2}\Big)\Big|\Big\};q\Big),$$
$$\hat{\sigma}^{\mathrm{M}}(\{z_t\}_{t=1}^T;q) := \hat{Q}\Big(\Big\{\Big|z_t - \hat{Q}\Big(\{z_s\}_{s=1}^T;\frac{1}{2}\Big)\Big|\Big\}_{t=1}^T;q\Big).$$

In the rest of the paper, we suppress the parameter q and write $\sigma^{\mathrm{M}}(Z)$ and $\hat{\sigma}^{\mathrm{M}}(\{z_t\}_{t=1}^T)$ for notational brevity. Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_T$ be a stationary sequence of random vectors, where $\boldsymbol{X}_t = (X_{t1},\ldots,X_{td})^{\mathsf{T}}$. As a generalization of MAD to the multivariate scenario, the population and sample MAD scatter matrices can be defined as

$$\mathbf{R}^{\mathrm{MAD}} := [\mathbf{R}_{jk}^{\mathrm{MAD}}] \text{ and } \hat{\mathbf{R}}^{\mathrm{MAD}} := [\hat{\mathbf{R}}_{jk}^{\mathrm{MAD}}],$$

where the entries of $\mathbf{R}^{\mathrm{MAD}}$ and $\hat{\mathbf{R}}^{\mathrm{MAD}}$ are given by

$$\mathbf{R}_{jj}^{\mathrm{MAD}} = \sigma^{\mathrm{M}}(X_{1j})^2, \quad \hat{\mathbf{R}}_{jj}^{\mathrm{MAD}} = \hat{\sigma}^{\mathrm{M}}(\{X_{tj}\}_{t=1}^T)^2,$$
$$\mathbf{R}_{jk}^{\mathrm{MAD}} = \frac{1}{4}\Big[\sigma^{\mathrm{M}}(X_{1j}+X_{1k})^2 - \sigma^{\mathrm{M}}(X_{1j}-X_{1k})^2\Big],$$
$$\hat{\mathbf{R}}_{jk}^{\mathrm{MAD}} = \frac{1}{4}\Big[\hat{\sigma}^{\mathrm{M}}(\{X_{tj}+X_{tk}\}_{t=1}^T)^2 - \hat{\sigma}^{\mathrm{M}}(\{X_{tj}-X_{tk}\}_{t=1}^T)^2\Big],$$

for $j \neq k \in \{1,\ldots,d\}$. In Han et al. (2014), $\mathbf{R}^{\mathrm{MAD}}$ and $\hat{\mathbf{R}}^{\mathrm{MAD}}$ have been studied under

[4] In Hampel (1974), q was set to 1/2 to achieve the best possible 50% breakdown point (i.e., the maximum proportion of outliers that the estimate can safely tolerate) and the most sharply bounded influence function (Hampel et al., 1986).

independent data.

For analyzing the consistency of the scatter matrix estimators, we introduce an identifiability condition on the distribution function of the random vector sequence.

Condition 4. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ be a stationary sequence of absolutely continuous random vectors. For any $j \neq k \in \{1, \ldots, d\}$, denote F_j, $\bar{F}_j$, $F^+_{j,k}$, $\bar{F}^+_{j,k}$, $F^-_{j,k}$, and $\bar{F}^-_{j,k}$ as the distribution functions of X_{1j}, $|X_{1j} - Q(X_{1j}; 1/2)|$, $X_{1j} + X_{1k}$, $|X_{1j} + X_{1k} - Q(X_{1j} + X_{1k}; 1/2)|$, $X_{1j} - X_{1k}$, and $|X_{1j} - X_{1k} - Q(X_{1j} - X_{1k}; 1/2)|$, respectively. We assume that the sequence $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ satisfies*

$$\inf_{|x - Q(F;q)| < \kappa_1} \frac{d}{dx} F(x) \geq \eta_1 \tag{4.5}$$

for any $F \in \{F_j, \bar{F}_j, F^+_{j,k}, \bar{F}^+_{j,k}, F^-_{j,k}, \bar{F}^-_{j,k} : j \neq k \in \{1, \ldots, d\}\}$ and some constants $\kappa_1, \eta_1 > 0$.

Condition 4 guarantees the identifiability of the medians of the distribution functions. This condition is standard in the literature on quantile statistics (Han et al., 2014; Belloni and Chernozhukov, 2011; Wang et al., 2012). Next, we present the rate of convergence for $\hat{\mathbf{R}}^{\mathrm{MAD}}$.

Theorem 22. *Under Conditions 3 and 4, for (T, d) large enough and any $\alpha \in (0, 1)$, with probability no smaller than $1 - 24\alpha^2$, we have*

$$\left\| \hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}} \right\|_{\max} \leq$$

$$\max\Big\{\frac{8}{\eta_1^2}\Big\{\sqrt{\frac{4D_1(\log d-\log\alpha)}{T}}+\frac{1}{T}\Big\}^2,\frac{8\sigma_{\max}^{\mathrm{M}}}{\eta_1}\Big\{\sqrt{\frac{4D_1(\log d-\log\alpha)}{T}}+\frac{1}{T}\Big\}\Big\}, \quad (4.6)$$

where $\sigma_{\max}^{\mathrm{M}} := \max\{\sigma^{\mathrm{M}}(X_j), \sigma^{\mathrm{M}}(X_j + X_k), \sigma^{\mathrm{M}}(X_j - X_k) : j \neq k \in \{1, \ldots, d\}\}$, $D_1 = 2^{a+5}K^2L_1(K^2 \vee 2)$, *and* η_1 *is defined in* (4.5).

The implications of Theorem 22 are as follows:

1. In the rates of convergence, the parameter $D_1 = 2^{a+5}K^2L_1(K^2 \vee 2)$ characterizes the effect of serial dependence on the consistency of the estimators. Specifically, in Condition 3, the degree of serial dependence in $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ is described by the parameters K, L_1 and a, which in turn modify the rates of convergence for $\hat{\mathbf{R}}^{\mathrm{MAD}}$ and $\hat{\mathbf{R}}_1^{\mathrm{MAD}}$ through D_1.

2. When D_1, η_1, $\sigma_{\max}^{\mathrm{M}}$ and $\tau_{\max}^{\mathrm{M}}$ are fixed, the rate of convergence for $\hat{\mathbf{R}}^{\mathrm{MAD}}$ reduces to $O_P(\sqrt{\log d/T})$. Han et al. (2014) derived similar rates of convergence for $\hat{\mathbf{R}}^{\mathrm{MAD}}$ under independent data points, and showed that the rate leads to optimal rates of convergence for various covariance estimators induced from $\hat{\mathbf{R}}^{\mathrm{MAD}}$.

3. Theorems 12, 15, 17, and 21 showed that the Kolmogorov dependence condition is satisfied under VAR models, α-mixing conditions, various covariance-based weak dependence conditions, and physical dependence conditions. Thus, Theorem 22 immediately implies consistency of $\hat{\mathbf{R}}^{\mathrm{MAD}}$ under these other dependence conditions.

The scatter matrix estimator $\hat{\mathbf{R}}^{\mathrm{MAD}}$ may not be positive semi-definite, while in many applications the estimand, $\mathbf{R}^{\mathrm{MAD}}$, is believed to be positive semi-definite. When a positive

semi-definite estimator is needed, we propose to project $\hat{\mathbf{R}}^{\mathrm{MAD}}$ into the cone of positive semi-definite matrices. Specifically, we define

$$\begin{aligned} &\tilde{\mathbf{R}}^{\mathrm{MAD}} = \arg\min_{\mathbf{R}} \left\| \hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R} \right\|_{\max}, \\ &\text{s.t. } \mathbf{R} \in S_\lambda := \{\mathbf{M} \in \mathbb{R}^{d\times d} : \mathbf{M}^{\mathsf{T}} = \mathbf{M}, \lambda_{\min}\mathbf{I}_d \preceq \mathbf{M} \preceq \lambda_{\max}\mathbf{I}_d\}, \end{aligned} \tag{4.7}$$

where $0 \le \lambda_{\min} < \lambda_{\max} \le \infty$ provides the lower and upper bounds of the eigenvalues of $\tilde{\mathbf{R}}^{\mathrm{MAD}}$. Problem (4.7) can be solved by a projection and contraction algorithm introduced in Xu and Shao (2012a). Appendix B.3 provides a brief summary of the algorithm[5]. The next theorem presents the rate of convergence for $\tilde{\mathbf{R}}^{\mathrm{MAD}}$ under the Kolmogorov dependence condition.

Theorem 23. *Under Conditions 3 and 4, if we assume $\mathbf{R}^{\mathrm{MAD}} \in S_\lambda$, then, for (T, d) large enough and any $\alpha \in (0, 1)$, with probability no smaller than $1 - 24\alpha^2$, we have*

$$\begin{aligned} &\left\| \tilde{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}} \right\|_{\max} \le \\ &\quad \max\Big\{ \frac{16}{\eta_1^2} \Big\{ \sqrt{\frac{4D_1(\log d - \log\alpha)}{T}} + \frac{1}{T} \Big\}^2, \frac{16\sigma^{\mathrm{M}}_{\max}}{\eta_1} \Big\{ \sqrt{\frac{4D_1(\log d - \log\alpha)}{T}} + \frac{1}{T} \Big\} \Big\}, \end{aligned} \tag{4.8}$$

where $\sigma^{\mathrm{M}}_{\max} := \max\{\sigma^{\mathrm{M}}(X_j), \sigma^{\mathrm{M}}(X_j + X_k), \sigma^{\mathrm{M}}(X_j - X_k) : j \ne k \in \{1, \dots, d\}\}$, $D_1 = 2^{a+5}K^2 L_1(K^2 \vee 2)$, and η_1 is defined in (4.5).

Theorem 23 shows that up to a constant, projecting $\hat{\mathbf{R}}^{\mathrm{MAD}}$ into the positive semi-definite cone doesn't lose rate of convergence, provided that $\mathbf{R}^{\mathrm{MAD}}$ is positive semi-definite.

[5]Replacing $\hat{\mathbf{R}}^{\mathrm{Q}}$ with $\hat{\mathbf{R}}^{\mathrm{MAD}}$, and $\tilde{\mathbf{R}}^{\mathrm{Q}}$ with $\tilde{\mathbf{R}}^{\mathrm{MAD}}$ in Appendix B.3 gives the algorithm for solving (4.7).

4.4 Proof of Main Results

In this section, we present the proofs of the main theorems. Proofs of the remaining results are collected in the appendix.

4.4.1 Proof of Results in Section 4.2

Proof of Theorem 12. Define $h(x) := I(x \leq b)$ and $h_\epsilon(x)$ be a smoothed version of h:

$$h_\epsilon(x) := \begin{cases} h(x), \text{ if } x < b - \epsilon \text{ or } x > b + \epsilon; \\ \frac{1}{4\epsilon^3}\{x^3 - 3bx^2 + 3(b^2 - \epsilon^2)x - b^3 + 3b\epsilon^2 + 2\epsilon^3\}, \text{ if } b - \epsilon \leq x \leq b + \epsilon. \end{cases} \tag{4.9}$$

where $\epsilon > 0$ is a constant that will be specified later. $h_\epsilon(x)$ is continuous with first order derivative

$$\frac{d}{dx}h_\epsilon(x) = \begin{cases} 0, \text{ if } x < b - \epsilon \text{ or } x > b + \epsilon; \\ \frac{3}{4\epsilon^3}\{(x - b)^2 - \epsilon^2\}, \text{ if } b - \epsilon \leq x \leq b + \epsilon. \end{cases}$$

Thus, $h_\epsilon(x)$ is Lipschitz continuous with $\mathrm{Lip}(h_\epsilon) = \sup_x |dh_\epsilon(x)/dx| = 3/(4\epsilon)$.

Next, we verify Condition 3. By the definition of covariance and the triangle inequality, we have

$$\Big|\mathbb{P}\Big(Y_t \leq b, \forall\, t \in \mathcal{S} \cup \mathcal{T}\Big) - \mathbb{P}\Big(Y_t \leq b, \forall\, t \in \mathcal{S}\Big)\mathbb{P}\Big(Y_{t'} \leq b, \forall\, t' \in \mathcal{T}\Big)\Big|$$

$$
\begin{aligned}
&= \Big| \mathrm{Cov}\Big\{ \prod_{t\in\mathcal{S}} h(Y_t), \prod_{t\in\mathcal{T}} h(Y_t) \Big\} \Big| \\
&\le \underbrace{\Big| \mathrm{Cov}\Big\{ \prod_{t\in\mathcal{S}} h(Y_t), \prod_{t\in\mathcal{T}} h(Y_t) \Big\} - \mathrm{Cov}\Big\{ \prod_{t\in\mathcal{S}} h_\epsilon(Y_t), \prod_{t\in\mathcal{T}} h_\epsilon(Y_t) \Big\} \Big|}_{A} + \\
&\qquad \underbrace{\Big| \mathrm{Cov}\Big\{ \prod_{t\in\mathcal{S}} h_\epsilon(Y_t), \prod_{t\in\mathcal{T}} h_\epsilon(Y_t) \Big\} \Big|}_{B}. \qquad (4.10)
\end{aligned}
$$

We first derive an upper bound for A. By the triangle inequality, we have

$$
\begin{aligned}
A \le & \Big| \mathrm{Cov}\Big\{ \prod_{t\in\mathcal{S}} h(Y_t), \prod_{t\in\mathcal{T}} h(Y_t) - \prod_{t\in\mathcal{T}} h_\epsilon(Y_t) \Big\} \Big| + \\
& \Big| \mathrm{Cov}\Big\{ \prod_{t\in\mathcal{S}} h(Y_t) - \prod_{t\in\mathcal{S}} h_\epsilon(Y_t), \prod_{t\in\mathcal{T}} h_\epsilon(Y_t) \Big\} \Big|. \qquad (4.11)
\end{aligned}
$$

For two random variables X and Y with $|X| \le 1$, we have

$$
|\,\mathrm{Cov}(X,Y)| = |\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y| \le \mathbb{E}|X||Y| + \mathbb{E}|X|\mathbb{E}|Y| \le 2\mathbb{E}|Y|. \qquad (4.12)
$$

Now, setting $X = \prod_{t\in\mathcal{S}} h(Y_t)$ and $Y = \prod_{t\in\mathcal{T}} h(Y_t) - \prod_{t\in\mathcal{T}} h_\epsilon(Y_t)$, we have

$$
\Big| \mathrm{Cov}\Big\{ \prod_{t\in\mathcal{S}} h(Y_t), \prod_{t\in\mathcal{T}} h(Y_t) - \prod_{t\in\mathcal{T}} h_\epsilon(Y_t) \Big\} \Big| \le 2\mathbb{E} \Big| \prod_{t\in\mathcal{T}} h(Y_t) - \prod_{t\in\mathcal{T}} h_\epsilon(Y_t) \Big|.
$$

Setting $X = \prod_{t\in\mathcal{T}} h_\epsilon(Y_t)$ and $Y = \prod_{t\in\mathcal{S}} h(Y_t) - \prod_{t\in\mathcal{S}} h_\epsilon(Y_t)$ in (4.12), we have

$$
\Big| \mathrm{Cov}\Big\{ \prod_{t\in\mathcal{S}} h(Y_t) - \prod_{t\in\mathcal{S}} h_\epsilon(Y_t), \prod_{t\in\mathcal{T}} h_\epsilon(Y_t) \Big\} \Big| \le 2\mathbb{E} \Big| \prod_{t\in\mathcal{S}} h(Y_t) - \prod_{t\in\mathcal{S}} h_\epsilon(Y_t) \Big|.
$$

Plugging the above two inequalities into (4.11), we have

$$\begin{aligned} A \leq & 2\mathbb{E}\Big|\prod_{t\in\mathcal{T}} h(Y_t) - \prod_{t\in\mathcal{T}} h_\epsilon(Y_t)\Big| + 2\mathbb{E}\Big|\prod_{t\in\mathcal{S}} h(Y_t) - \prod_{t\in\mathcal{S}} h_\epsilon(Y_t)\Big| \\ \leq & 2(|\mathcal{S}| + |\mathcal{T}|)\mathbb{E}\Big|h(Y_t) - h_\epsilon(Y_t)\Big|. \end{aligned}$$

The last inequality is due to the fact that

$$\Big|\prod_{t=1}^{m} a_t - \prod_{t=1}^{m} b_t\Big| \leq \sum_{t=1}^{m} |a_t - b_t| \tag{4.13}$$

for $0 \leq a_t, b_t \leq 1$. Noting that $|h(Y_t) - h_\epsilon(Y_t)| \leq 1$ and $h(Y_t) - h_\epsilon(Y_t)$ is non-zero only when $b - \epsilon \leq Y_t \leq b + \epsilon$, using Assumption 3, we have

$$A \leq 2(|\mathcal{S}| + |\mathcal{T}|)\mathbb{P}(b - \epsilon \leq Y_t \leq b + \epsilon) \leq 4H(|\mathcal{S}| + |\mathcal{T}|)\epsilon. \tag{4.14}$$

Now we derive the upper bound of B. Since $\|\mathbf{A}\|_2 < 1$, the process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ has moving average representation $\boldsymbol{X}_t = \sum_{\ell=0}^{\infty} \mathbf{A}^\ell \boldsymbol{\epsilon}_{t-\ell}$. Define $\boldsymbol{X}_t^{[p]}$ to be a finite order moving average process: $\boldsymbol{X}_t^{[p]} := \sum_{\ell=0}^{p-1} \mathbf{A}^\ell \boldsymbol{\epsilon}_{t-\ell}$ where $p = d(\mathcal{S}, \mathcal{T})$. Now, depending on the choice of Y_t, we define

$$\mathbf{e} := \begin{cases} \mathbf{e}_j, & \text{if } Y_t = X_{tj}; \\ \mathbf{e}_j + \mathbf{e}_k, & \text{if } Y_t = X_{tj} + X_{tk}; \\ \mathbf{e}_j - \mathbf{e}_k, & \text{if } Y_t = X_{tj} - X_{tk}, \end{cases} \tag{4.15}$$

so that $Y_t = \mathbf{e}^\mathsf{T} \boldsymbol{X}_t$. Here $\mathbf{e}_j$ and $\mathbf{e}_k$ are the j-th and k-th columns of the identity matrix. Define $Y_t^{[p]} := \mathbf{e}^\mathsf{T} \boldsymbol{X}_t^{[p]}$. Using $Y_t^{[p]}$, we can upper bound B in (4.10) by

$$B \leq \underbrace{\left|\mathrm{Cov}\Big\{\prod_{t\in\mathcal{S}} h_\epsilon(Y_t) - \prod_{t\in\mathcal{S}} h_\epsilon(Y_t^{[p]}), \prod_{t\in\mathcal{T}} h_\epsilon(Y_t)\Big\}\right|}_{B_1} + \underbrace{\left|\mathrm{Cov}\Big\{\prod_{t\in\mathcal{S}} h_\epsilon(Y_t^{[p]}), \prod_{t\in\mathcal{T}} h_\epsilon(Y_t) - \prod_{t\in\mathcal{T}} h_\epsilon(Y_t^{[p]})\Big\}\right|}_{B_2}$$
$$+ \underbrace{\left|\mathrm{Cov}\Big\{\prod_{t\in\mathcal{S}} h_\epsilon(Y_t^{[p]}), \prod_{t\in\mathcal{T}} h_\epsilon(Y_t^{[p]})\Big\}\right|}_{B_3}. \tag{4.16}$$

Note that $\{Y_t^{[p]} : t \in \mathcal{S}\}$ only depends on $\{\boldsymbol{\epsilon}_t : \min(\mathcal{S}) - p < t \leq \max(\mathcal{S})\}$ and $\{Y_t^{[p]} : t \in \mathcal{T}\}$ only depends on $\{\boldsymbol{\epsilon}_t : \min(\mathcal{T}) - p < t \leq \max(\mathcal{T})\}$. Since $p = d(\mathcal{S}, \mathcal{T}) = \min(\mathcal{T}) - \max(\mathcal{S})$, we have that $\prod_{t\in\mathcal{S}} h_\epsilon(Y_t^{[p]})$ and $\prod_{t\in\mathcal{T}} h_\epsilon(Y_t^{[p]})$ are independent. Thus, we have $B_3 = 0$. Regarding B_1, using (4.12) and (4.13), we have

$$\begin{aligned} B_1 \leq & 2\mathbb{E}\Big|\prod_{t\in\mathcal{S}} h_\epsilon(Y_t) - \prod_{t\in\mathcal{S}} h_\epsilon(Y_t^{[p]})\Big| \leq 2|\mathcal{S}|\mathbb{E}|h_\epsilon(Y_t) - h_\epsilon(Y_t^{[p]})| \\ \leq & 2|\mathcal{S}|\mathrm{Lip}(h_\epsilon)\mathbb{E}|Y_t - Y_t^{[p]}|. \end{aligned} \tag{4.17}$$

Plugging in $\mathrm{Lip}(h_\epsilon) = 3/(4\epsilon)$, $Y_t = \mathbf{e}^\mathsf{T}\boldsymbol{X}_t = \sum_{\ell=0}^\infty \mathbf{e}^\mathsf{T}\mathbf{A}^\ell\boldsymbol{\epsilon}_{t-\ell}$ and $Y_t^{[p]} = \mathbf{e}^\mathsf{T}\boldsymbol{X}_t^{[p]} = \sum_{\ell=0}^{p-1} \mathbf{e}^\mathsf{T}\mathbf{A}^\ell\boldsymbol{\epsilon}_{t-\ell}$, we obtain

$$B_1 \leq \frac{3}{2\epsilon}|\mathcal{S}|\mathbb{E}\Big|\sum_{\ell=p}^\infty \mathbf{e}^\mathsf{T}\mathbf{A}^\ell\boldsymbol{\epsilon}_{t-\ell}\Big| \leq \frac{3}{2\epsilon}|\mathcal{S}|\sum_{\ell=p}^\infty \mathbb{E}\Big|\mathbf{e}^\mathsf{T}\mathbf{A}^\ell\boldsymbol{\epsilon}_{t-\ell}\Big| \leq \frac{3C|\mathcal{S}|\|\mathbf{A}\|_2^p}{\epsilon(1-\|\mathbf{A}\|_2)}.$$

The last inequality is due to Assumptions 1 and 2 on the VAR process. Applying similar

arguments to B_2, we have $B_2 \leq 3C|\mathcal{T}|\|\mathbf{A}\|_2^p/\{\epsilon(1-\|\mathbf{A}\|_2)\}$. Thus, we have

$$B \leq B_1 + B_2 \leq \frac{3C(|\mathcal{S}|+|\mathcal{T}|)\|\mathbf{A}\|_2^p}{\epsilon(1-\|\mathbf{A}\|_2)}. \tag{4.18}$$

Combining (4.10), (4.14), and (4.18), we have

$$\begin{aligned}&\Big|\mathbb{P}\Big(Y_t \leq b, \forall\, t \in \mathcal{S}\cup\mathcal{T}\Big) - \mathbb{P}\Big(Y_t \leq b, \forall\, t\in\mathcal{S}\Big)\mathbb{P}\Big(Y_{t'} \leq b, \forall\, t'\in\mathcal{T}\Big)\Big| \\ &\leq (|\mathcal{S}|+|\mathcal{T}|)\Big\{4H\epsilon + \frac{3C\|\mathbf{A}\|_2^p}{\epsilon(1-\|\mathbf{A}\|_2)}\Big\}.\end{aligned}$$

Now setting $\epsilon = \|\mathbf{A}\|_2^{p/2}$, we have

$$\begin{aligned}&\Big|\mathbb{P}\Big(Y_t \leq b, \forall\, t \in \mathcal{S}\cup\mathcal{T}\Big) - \mathbb{P}\Big(Y_t \leq b, \forall\, t\in\mathcal{S}\Big)\mathbb{P}\Big(Y_{t'} \leq b, \forall\, t'\in\mathcal{T}\Big)\Big| \\ &\leq \Big(4H + \frac{3C}{1-\|\mathbf{A}\|_2}\Big)(|\mathcal{S}|+|\mathcal{T}|)\|\mathbf{A}\|_2^{p/2}.\end{aligned}$$

To verify (4.1), we note that for any $k \leq 0$,

$$\begin{aligned}\sum_{s=0}^{\infty}(s+1)^k\|\mathbf{A}\|_2^{s/2} &\leq \sum_{s=0}^{\infty}(s+1)\cdots(s+k)\|\mathbf{A}\|_2^{s/2} \\ &= \frac{d^k}{dx^k}\Big(\frac{1}{1-x}\Big)\Big|_{x=\sqrt{\|\mathbf{A}\|_2}} = \frac{k!}{(1-\sqrt{\|\mathbf{A}\|_2})^{k+1}}.\end{aligned}$$

Thus, Condition 3 is satisfied with $K^2 = 4H + 3C/(1-\|\mathbf{A}\|_2)$, $\Psi(u,v) = u+v$, $L_1 = L = 1/(1-\sqrt{\|\mathbf{A}\|_2})$, and $a = 1$. This completes the proof. □

Proof of Theorem 15. Since $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is α-mixing, $\{Y_t\}_{t\in\mathbb{Z}}$ is also α-mixing. By the defi-

nition of α-mixing coefficient, we have

$$\Big|\mathbb{P}\Big(Y_t \le b, \forall\, t \in \mathcal{S} \cup \mathcal{T}\Big) - \mathbb{P}\Big(Y_t \le b, \forall\, t \in \mathcal{S}\Big)\mathbb{P}\Big(Y_{t'} \le b, \forall\, t' \in \mathcal{T}\Big)\Big| \le \boldsymbol{\alpha}\Big\{d(\mathcal{S}, \mathcal{T})\Big\}.$$

Verification of (4.1) follows the proof of Proposition 8 in Doukhan and Neumann (2007), and is omitted here. □

Proof of Theorem 17. Let h_ϵ be defined in (4.9). Using the same arguments as in the proof of Theorem 12, we still have (4.10) and (4.14). It remains to derive an upper bound for B. Since $\mathrm{Lip}(h_\epsilon) = 3/(4\epsilon)$, it's easy to check that for any $\mathbf{x}_1, \ldots, \mathbf{x}_u, \mathbf{y}_1, \ldots, \mathbf{y}_u \in \mathbb{R}^d$, we have

$$\Big|\prod_{t=1}^{u} h_\epsilon(\mathbf{e}^\mathsf{T}\mathbf{x}_t) - \prod_{t=1}^{u} h_\epsilon(\mathbf{e}^\mathsf{T}\mathbf{y}_t)\Big| \le \frac{3}{4\epsilon}\sum_{t=1}^{u} |\mathbf{e}^\mathsf{T}(\mathbf{x}_t - \mathbf{y}_t)| \le \frac{3}{2\epsilon}\sum_{t=1}^{u} \|\mathbf{x}_t - \mathbf{y}_t\|_q,$$

for any $0 < q \le \infty$, where $\mathbf{e}$ is defined in (4.15). This implies that the function $g(\mathbf{x}_1, \ldots, \mathbf{x}_u)$ $= \prod_{t=1}^{u} h_\epsilon(\mathbf{e}^\mathsf{T}\mathbf{x}_t)$ is Lipschitz with $\mathrm{Lip}(g) \le 3/(2\epsilon)$. Thus, by the assumption that $\{\boldsymbol{X}_t\}_{t\in\boldsymbol{Z}}$ is $(\Lambda^{(1)}, \psi, \zeta)$-weakly dependent, we have

$$B = \Big|\mathrm{Cov}\Big\{g(\{\boldsymbol{X}_t : t \in \mathcal{S}\}), g(\{\boldsymbol{X}_t : t \in \mathcal{T}\})\Big\}\Big| \le \psi(\mathrm{Lip}(g), \mathrm{Lip}(g), |\mathcal{S}|, |\mathcal{T}|)\zeta\{d(\mathcal{S}, \mathcal{T})\}.$$

Combining the above upper bound with (4.14), we have

$A+B \le$

$$\begin{cases} 4H(|\mathcal{S}|+|\mathcal{T}|)\epsilon + \frac{3}{2\epsilon}|\mathcal{T}|\zeta\{d(\mathcal{S},\mathcal{T})\}, \text{ if } \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}} \text{ is } \theta\text{-dependent;} \\ (|\mathcal{S}|+|\mathcal{T}|)\Big\{4H\epsilon + \frac{3}{2\epsilon}\zeta\{d(\mathcal{S},\mathcal{T})\}\Big\}, \text{ if } \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}} \text{ is } \eta\text{-dependent;} \\ 4H(|\mathcal{S}|+|\mathcal{T}|)\epsilon + \frac{9}{4\epsilon^2}|\mathcal{S}||\mathcal{T}|\zeta\{d(\mathcal{S},\mathcal{T})\}, \text{ if } \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}} \text{ is } \kappa\text{-dependent;} \\ 4H(|\mathcal{S}|+|\mathcal{T}|)\epsilon + \Big\{\frac{3}{2\epsilon}(|\mathcal{S}|+|\mathcal{T}|) + \frac{9}{4\epsilon^2}|\mathcal{S}||\mathcal{T}|\Big\}\zeta\{d(\mathcal{S},\mathcal{T})\}, \text{ if } \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}} \text{ is } \lambda\text{-dependent.} \end{cases}$$

Thus, if $\{\boldsymbol{X}_t\}_{t\in\boldsymbol{Z}}$ is θ- or η-dependent, setting $\epsilon = \sqrt{\zeta\{d(\mathcal{S},\mathcal{T})\}}$ gives the desired result. If $\{\boldsymbol{X}_t\}_{t\in\boldsymbol{Z}}$ is κ-dependent, setting $\epsilon = \zeta\{d(\mathcal{S},\mathcal{T})\}^{1/3}$ gives the desired result. If $\{\boldsymbol{X}_t\}_{t\in\boldsymbol{Z}}$ is λ-dependent, without loss of generality, we may assume that $\zeta\{d(\mathcal{S},\mathcal{T})\} \leq 1$. Thus, we have

$$A + B \leq \Big\{4H\epsilon + \frac{3}{2\epsilon^2}\zeta\{d(\mathcal{S},\mathcal{T})\}\Big\}(|\mathcal{S}|+|\mathcal{T}|) + \frac{9}{4\epsilon^2}|\mathcal{S}||\mathcal{T}|\zeta\{d(\mathcal{S},\mathcal{T})\}.$$

Setting $\epsilon = \zeta\{d(\mathcal{S},T)\}^{1/3}$ gives the desired result. □

Proof of Theorem 21. Let h_ϵ and $\mathbf{e}$ be defined in (4.9) and (4.15). Using the same arguments as in the proof of Theorem 12, we still have (4.10) and (4.14). To derive an upper bound on B, let $\{\boldsymbol{\epsilon}'_t\}_{t\in\boldsymbol{Z}}$ and $\{\boldsymbol{\epsilon}''_t\}_{t\in\boldsymbol{Z}}$ be two i.i.d. copies of $\{\boldsymbol{\epsilon}_t\}_{t\in\boldsymbol{Z}}$. Let $p = d(\mathcal{S},\mathcal{T})$. Define $J(t,p) := \{t-p, t-p-1, t-p-2, \ldots\}$ and

$$\mathcal{G}_t := (\ldots, \boldsymbol{\epsilon}'_{t-p-1}, \boldsymbol{\epsilon}'_{t-p}, \boldsymbol{\epsilon}_{t-p+1}, \ldots, \boldsymbol{\epsilon}_t),$$

$$\mathcal{H}_t := (\ldots, \boldsymbol{\epsilon}''_{t-p-1}, \boldsymbol{\epsilon}''_{t-p}, \boldsymbol{\epsilon}_{t-p+1}, \ldots, \boldsymbol{\epsilon}_t).$$

$\mathcal{G}_t$ and $\mathcal{H}_t$ are coupled versions of $\mathcal{F}_t$ with $\boldsymbol{\epsilon}_j$ replaced by $\boldsymbol{\epsilon}'_j$ and $\boldsymbol{\epsilon}''_j$ if $j \in J(t,p)$. Now define the process $\{\boldsymbol{X}_t^{[p]}\}_{t\in\boldsymbol{Z}}$ by

$$\boldsymbol{X}_t^{[p]} := \begin{cases} \boldsymbol{g}(\mathcal{G}_t) \text{ if } t \in \mathcal{S}; \\ \boldsymbol{g}(\mathcal{H}_t) \text{ if } t \in \mathcal{T}, \end{cases}$$

and $Y_t^{[p]} = \mathbf{e}^\mathsf{T}\boldsymbol{X}_t^{[p]}$. For the same reason as in the proof of Theorem 12, B in (4.10) can be upper bounded by (4.16). Note that by the definition of $\boldsymbol{X}_t^{[p]}$, $\{\boldsymbol{X}_t^{[p]} : t \in \mathcal{S}\}$ and $\{\boldsymbol{X}_t^{[p]} : t \in \mathcal{T}\}$ are independent. Thus, we still have $B_3 = 0$. Using the same technique as in (4.17), we have

$$B_1 \le \frac{3}{2\epsilon}|\mathcal{S}|\mathbb{E}\Big|Y_t - Y_t^{[p]}\Big| \le \frac{3}{\epsilon}|\mathcal{S}| \max_{j=1,\dots,d}\Big(\mathbb{E}|X_{tj} - X_{tj}^{[p]}|\Big) \le \frac{3}{\epsilon}|\mathcal{S}| \max_{j=1,\dots,d}\theta_{p,j},$$

where the last equality is due to stationarity. Using similar arguments, we can also obtain $B_2 \le 3|\mathcal{T}| \max_{j=1,\dots,d}\theta_{p,j}/\epsilon$. Thus, we have

$$B \le B_1 + B_2 \le \frac{3}{\epsilon}(|\mathcal{S}| + |\mathcal{T}|) \max_{j=1,\dots,d}\theta_{p,j}.$$

Combining the above inequality with (4.14), we have

$$A + B \le (|\mathcal{S}| + |\mathcal{T}|)\Big(4H\epsilon + \frac{3}{\epsilon}\max_{j=1,\dots,d}\theta_{p,j}\Big).$$

Setting $\epsilon = \max_{j=1,\dots,d}\sqrt{\theta_{p,j}}$ completes the proof. □

4.4.2 Proof of Results in Section 4.3

Proof of Theorem 22. Equation (4.5) implies that

$$F\Big\{F^{-1}\Big(\frac{1}{2}\Big)+\frac{t}{2}\Big\}-\frac{1}{2}=F\Big\{F^{-1}\Big(\frac{1}{2}\Big)+\frac{t}{2}\Big\}-F\Big\{F^{-1}\Big(\frac{1}{2}\Big)\Big\}\geq\frac{\eta_1 t}{2},$$
$$\frac{1}{2}-F\Big\{F^{-1}\Big(\frac{1}{2}\Big)-\frac{t}{2}\Big\}=F\Big\{F^{-1}\Big(\frac{1}{2}\Big)\Big\}-F\Big\{F^{-1}\Big(\frac{1}{2}\Big)-\frac{t}{2}\Big\}\geq\frac{\eta_1 t}{2},$$

for $0<t/2\leq\kappa$ and $F\in\{F_j,\bar{F}_j:j=1,\ldots,d\}$. We allow D_2 in (C.13) to depend on T. Specifically, we define

$$D_{2,T}=2\Big\{2L(T,d)(K^2\vee 2)\Big\}^{1/(a+2)}, \tag{4.19}$$

and correspondingly, let

$$\varphi_T(x):=\frac{Tx^2}{D_1+D_{2,T}T^{(a+1)/(a+2)}x^{(2a+3)/(a+2)}},\ \text{ for }\ x>0. \tag{4.20}$$

It's easy to check that φ_T is non-decreasing on $(0,\infty)$ by investigating the derivative of $\log\varphi_T(x)$. Thus, using Lemma 13, we have, for any $j\in\{1,\ldots,d\}$,

$$\begin{aligned}\mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}(\{X_{tj}\}_{t=1}^T)-\sigma^{\mathrm{M}}(X_j)\Big|>t\Big\}\leq&3\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{2}-\frac{1}{T}\Big)\Big\}+3\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{2}\Big)\Big\}\\ \leq&6\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{2}-\frac{1}{T}\Big)\Big\},\end{aligned} \tag{4.21}$$

when $0 < t/2 < \kappa$ and $\eta_1 t/2 > 1/T$. Now, by the definitions of $\hat{\mathbf{R}}_{jj}^{\mathrm{MAD}}$ and $\mathbf{R}_{jj}^{\mathrm{MAD}}$, we have

$$\begin{aligned}
&\mathbb{P}\Big(|\hat{\mathbf{R}}_{jj}^{\mathrm{MAD}} - \mathbf{R}_{jj}^{\mathrm{MAD}}| > t\Big) \\
=&\mathbb{P}\Big[\Big|\hat{\sigma}^{\mathrm{M}}(\{X_{tj}\}_{t=1}^{T})^2 - \sigma^{\mathrm{M}}(X_j)^2\Big| > t\Big] \\
\leq&\mathbb{P}\Big[\Big\{\hat{\sigma}^{\mathrm{M}}(\{X_{tj}\}_{t=1}^{T}) - \sigma^{\mathrm{M}}(X_j)\Big\}^2 + 2\Big|\sigma^{\mathrm{M}}(X_j)\Big\{\hat{\sigma}^{\mathrm{M}}(\{X_{tj}\}_{t=1}^{T}) - \sigma^{\mathrm{M}}(X_j)\Big\}\Big| > t\Big] \\
\leq&\mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}(\{X_{tj}\}_{t=1}^{T}) - \sigma^{\mathrm{M}}(X_j)\Big| > \sqrt{\frac{t}{2}}\Big\} + \\
&\quad\mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}(\{X_{tj}\}_{t=1}^{T}) - \sigma^{\mathrm{M}}(X_j)\Big| > \frac{t}{4\sigma^{\mathrm{M}}(X_j)}\Big\}. \qquad (4.22)
\end{aligned}$$

Applying (4.21), we have

$$\begin{aligned}
&\mathbb{P}\Big(|\hat{\mathbf{R}}_{jj}^{\mathrm{MAD}} - \mathbf{R}_{jj}^{\mathrm{MAD}}| > t\Big) \\
&\leq 6\exp\Big\{-\varphi_T\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}} - \frac{1}{T}\Big)\Big\} + 6\exp\Big[-\varphi_T\Big\{\frac{\eta_1 t}{8\sigma^{\mathrm{M}}(X_j)} - \frac{1}{T}\Big\}\Big] \\
&\leq 12\max\Big\{\exp\Big\{-\varphi_T\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}} - \frac{1}{T}\Big)\Big\}, \exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{8\sigma_{\max}^{\mathrm{M}}} - \frac{1}{T}\Big)\Big\}\Big\}. \qquad (4.23)
\end{aligned}$$

Next, we derive the concentration inequality about $\hat{\mathbf{R}}_{jk}^{\mathrm{MAD}}$ for $j \neq k$. Again, using Lemma 13, we have, for $j \neq k$,

$$\mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}(\{X_{tj}+X_{tk}\}_{t=1}^{T}) - \sigma^{\mathrm{M}}(X_j+X_k)\Big| > t\Big\} \leq 6\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{2} - \frac{1}{T}\Big)\Big\}, \qquad (4.24)$$

$$\mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}(\{X_{tj}-X_{tk}\}_{t=1}^{T}) - \sigma^{\mathrm{M}}(X_j-X_k)\Big| > t\Big\} \leq 6\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{2} - \frac{1}{T}\Big)\Big\}. \qquad (4.25)$$

By the definitions of $\hat{\mathbf{R}}_{jk}^{\mathrm{MAD}}$ and $\mathbf{R}_{jk}^{\mathrm{MAD}}$, we have

$$\begin{aligned}
&\mathbb{P}\Big(|\hat{\mathbf{R}}_{jk}^{\mathrm{MAD}} - \mathbf{R}_{jk}^{\mathrm{MAD}}| > t\Big) \\
=&\Big(\Big[\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}+X_{tk}\}_{t=1}^{T}\Big)^2 - \sigma^{\mathrm{M}}(X_j+X_k)^2\Big] + \\
&\qquad \Big[\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}-X_{tk}\}_{t=1}^{T}\Big)^2 - \sigma^{\mathrm{M}}(X_j-X_k)^2\Big] > 4t\Big) \\
\leq& \underbrace{\mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}+X_{tk}\}_{t=1}^{T}\Big)^2 - \sigma^{\mathrm{M}}(X_j+X_k)^2\Big| > 2t\Big\}}_{P_1} + \\
&\qquad \underbrace{\mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}-X_{tk}\}_{t=1}^{T}\Big)^2 - \sigma^{\mathrm{M}}(X_j-X_k)^2\Big| > 2t\Big\}}_{P_2}. \qquad (4.26)
\end{aligned}$$

Using the same technique as in (4.22), we have

$$\begin{aligned}
P_1 \leq& \mathbb{P}\Big[\Big\{\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}+X_{tk}\}_{t=1}^{T}\Big) - \sigma^{\mathrm{M}}(X_j+X_k)\Big\}^2 + \\
&\qquad 2\Big|\sigma^{\mathrm{M}}(X_j+X_k)\Big\{\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}+X_{tk}\}_{t=1}^{T}\Big) - \sigma^{\mathrm{M}}(X_j+X_k)\Big\}\Big| > 2t\Big] \\
\leq& \mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}+X_{tk}\}_{t=1}^{T}\Big) - \sigma^{\mathrm{M}}(X_j+X_k)\Big| > \sqrt{t}\Big\} + \\
&\qquad \mathbb{P}\Big[\Big|\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}+X_{tk}\}_{t=1}^{T}\Big) - \sigma^{\mathrm{M}}(X_j+X_k)\Big| > \frac{t}{2\sigma^{\mathrm{M}}(X_j+X_k)}\Big], \qquad (4.27)
\end{aligned}$$

and similarly

$$\begin{aligned}
P_2 \leq& \mathbb{P}\Big[\Big\{\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}-X_{tk}\}_{t=1}^{T}\Big) - \sigma^{\mathrm{M}}(X_j-X_k)\Big\}^2 + \\
&\qquad 2\Big|\sigma^{\mathrm{M}}(X_j-X_k)\Big\{\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}-X_{tk}\}_{t=1}^{T}\Big) - \sigma^{\mathrm{M}}(X_j-X_k)\Big\}\Big| > 2t\Big] \\
\leq& \mathbb{P}\Big\{\Big|\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}-X_{tk}\}_{t=1}^{T}\Big) - \sigma^{\mathrm{M}}(X_j-X_k)\Big| > \sqrt{t}\Big\} +
\end{aligned}$$

$$\mathbb{P}\Big[\Big|\hat{\sigma}^{\mathrm{M}}\Big(\{X_{tj}-X_{tk}\}_{t=1}^{T}\Big)-\sigma^{\mathrm{M}}(X_j-X_k)\Big|>\frac{t}{2\sigma^{\mathrm{M}}(X_j-X_k)}\Big]. \tag{4.28}$$

Applying (4.24) and (4.25) to the above two inequalities and noting that $\sigma^{\mathrm{M}}(X_j+X_k)\leq \sigma^{\mathrm{M}}_{\max}$, $\sigma^{\mathrm{M}}(X_j-X_k)\leq\sigma^{\mathrm{M}}_{\max}$, we obtain

$$\begin{aligned}
P_1 \leq& 6\exp\Big\{-\varphi_T\Big(\frac{\eta_1\sqrt{t}}{2}-\frac{1}{T}\Big)\Big\}+6\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{4\sigma^{\mathrm{M}}_{\max}}-\frac{1}{T}\Big)\Big\},\\
P_2 \leq& 6\exp\Big\{-\varphi_T\Big(\frac{\eta_1\sqrt{t}}{2}-\frac{1}{T}\Big)\Big\}+6\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{4\sigma^{\mathrm{M}}_{\max}}-\frac{1}{T}\Big)\Big\}.
\end{aligned}$$

Plugging the above two inequalities into (4.26), we have

$$\begin{aligned}
&\mathbb{P}\Big(|\hat{\mathbf{R}}^{\mathrm{MAD}}_{jk}-\mathbf{R}^{\mathrm{MAD}}_{jk}|>t\Big)\\
\leq& 12\exp\Big\{-\varphi_T\Big(\frac{\eta_1\sqrt{t}}{2}-\frac{1}{T}\Big)\Big\}+12\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{8\sigma^{\mathrm{M}}_{\max}}-\frac{1}{T}\Big)\Big\}\\
\leq& 24\max\Big\{\exp\Big\{-\varphi_T\Big(\frac{\eta_1\sqrt{t}}{2}-\frac{1}{T}\Big)\Big\},\exp\Big\{-\varphi_T\Big(\frac{\eta_1 t}{8\sigma^{\mathrm{M}}_{\max}}-\frac{1}{T}\Big)\Big\}\Big\}.
\end{aligned} \tag{4.29}$$

Combining (4.23) and (4.29), we have

$$\begin{aligned}
&\mathbb{P}\Big(\Big\|\hat{\mathbf{R}}^{\mathrm{MAD}}-\mathbf{R}^{\mathrm{MAD}}\Big\|_{\max}>t\Big)\\
\leq& 24\max\Big\{\exp\Big\{2\log d-\varphi_T\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}}-\frac{1}{T}\Big)\Big\},\exp\Big\{2\log d-\varphi_T\Big(\frac{\eta_1 t}{8\sigma^{\mathrm{M}}_{\max}}-\frac{1}{T}\Big)\Big\}\Big\}.
\end{aligned} \tag{4.30}$$

Next, we simplify the above concentration bound using the special structure of function φ_T. Let

$$b_1(t) := \exp\Big\{2\log d - \varphi_T\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}} - \frac{1}{T}\Big)\Big\} \text{ and } b_2(t) := \exp\Big\{2\log d - \varphi_T\Big(\frac{\eta_1 t}{8\sigma_{\max}^{\mathrm{M}}} - \frac{1}{T}\Big)\Big\}.$$

We discuss the form of the concentration bound in two scenarios:

(i) If $b_1(t) \geq b_2(t)$, we focus on $b_1(t)$. We remind that by the definition of function φ_T, we have

$$\varphi_T\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}} - \frac{1}{T}\Big) = \frac{T\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}} - \frac{1}{T}\Big)^2}{D_1 + D_{2,T}T^{(a+1)/(a+2)}\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}} - \frac{1}{T}\Big)^{(2a+3)/(a+2)}},$$

where D_1 and $D_{2,T}$ are defined in (C.14) and (4.19). To simplify the denominator on the right-hand side of the above equation, we require that

$$D_1 \geq D_{2,T}T^{(a+1)/(a+2)}\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}} - \frac{1}{T}\Big)^{(2a+3)/(a+2)}. \tag{4.31}$$

Then we have $\varphi_T\{\eta_1\sqrt{t}/(2\sqrt{2}) - 1/T\} \geq T\{\eta_1\sqrt{t}/(2\sqrt{2}) - 1/T\}^2/(2D_1)$. By the definition of $b_1(t)$, we have

$$b_1(t) \leq \exp\Big\{2\log d - \frac{T}{2D_1}\Big(\frac{\eta_1}{2}\sqrt{\frac{t}{2}} - \frac{1}{T}\Big)^2\Big\}.$$

Setting $\exp\Big[2\log d - T\Big\{\eta_1\sqrt{t}/(2\sqrt{2}) - 1/T\Big\}^2/(2D_1)\Big] = \alpha^2$ for some $\alpha \in (0,1)$,

we obtain

$$t = \frac{8}{\eta_1^2}\Big\{\sqrt{\frac{4D_1(\log d - \log \alpha)}{T}} + \frac{1}{T}\Big\}^2 := t_1(T, d). \tag{4.32}$$

Under (4.32), for $d > 1/\alpha$, we have $\eta_1\sqrt{t}/(2\sqrt{2}) - 1/T \leq \sqrt{8D_1 \log d/T}$. Thus, (4.31) holds if we require

$$D_1 \geq D_{2,T} T^{(a+1)/(a+2)} (8D_1 \log d/T)^{(a+3/2)/(a+2)}.$$

Using the definitions of D_1 and $D_{2,T}$, it follows that (4.31) holds when we have

$$L(T, d) \leq \frac{K\sqrt{L_1}}{2^{7a/2+5}\sqrt{K^2 \vee 2}} \frac{\sqrt{T}}{(\log d)^{a+3/2}}. \tag{4.33}$$

Thus, (4.31) is guaranteed by (4.2) in Condition 3.

(ii) If $b_1(t) < b_2(t)$, we follow a similar argument as in (i) and require that

$$D_1 \geq D_{2,T} T^{(a+1)/(a+2)} \Big(\frac{\eta_1 t}{8\sigma_{\max}^{\mathrm{M}}} - \frac{1}{T}\Big)^{(2a+3)/(a+2)}. \tag{4.34}$$

This leads to $\varphi_T\{\eta_1 t/(8\sigma_{\max}^{\mathrm{M}}) - 1/T\} \geq T\{\eta_1 t/(8\sigma_{\max}^{\mathrm{M}}) - 1/T\}^2/(2D_1)$. By the definition of $b_2(t)$, we have

$$b_2(t) \leq \exp\Big\{2\log d - \frac{T}{2D_1}\Big(\frac{\eta_1 t}{8\sigma_{\max}^{\mathrm{M}}} - \frac{1}{T}\Big)^2\Big\}.$$

Setting $\exp\Big[2\log d - T\Big\{\eta_1 t/(8\sigma^{\mathrm{M}}_{\max}) - 1/T\Big\}^2/(2D_1)\Big] = \alpha^2$, we obtain

$$t = \frac{8\sigma^{\mathrm{M}}_{\max}}{\eta_1}\Big\{\sqrt{\frac{4D_1(\log d - \log\alpha)}{T}} + \frac{1}{T}\Big\} := t_2(T,d). \tag{4.35}$$

Under (4.35), we have $\eta_1 t/(8\sigma^{\mathrm{M}}_{\max}) - 1/T \leq \sqrt{8D_1\log d/T}$ if $d > 1/\alpha$. Thus, (4.34) holds if we again require

$$D_1 \geq D_{2,T}T^{(a+1)/(a+2)}(8D_1\log d/T)^{(a+3/2)/(a+2)}.$$

Now, using the definitions of D_1 and $D_{2,T}$, we obtain that (4.34) is also guaranteed by (4.2) in Condition 3.

Now we summarize the discussion above and derive the final rate of convergence. In (4.30), we set $t = \max\Big\{t_1(T,d), t_2(T,d)\Big\}$ and require that (4.33) holds. When $t_1(T,d) \geq t_2(T,d)$, we have $t = t_1(T,d)$. Thus, together with (4.33), we have $b_1(t) \leq \alpha^2$. Since $b_2(t)$ is nonincreasing in t, we have $b_2\{t_1(T,d)\} \leq b_2\{t_2(T,d)\} \leq \alpha^2$. The last inequality is ensured by (4.33). Thus, we obtain

$$\mathbb{P}\Big(\Big\|\hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}}\Big\|_{\max} > t\Big) \leq 24\max\Big\{b_1(t), b_2(t)\Big\} \leq 24\alpha^2. \tag{4.36}$$

On the other hand, when $t_1(T,d) < t_2(T,d)$, we have $t = t_2(T,d)$. Thus, together with (4.33), we have $b_2(t) \leq \alpha^2$. Since $b_1(t)$ is nonincreasing in t, we have $b_1\{t_2(T,d)\} \leq b_1\{t_1(T,d)\} \leq \alpha^2$, where the last inequality is ensured by (4.33). Thus, again, we can

obtain (4.36). So, in either case, we have

$$\mathbb{P}\Big(\Big\|\hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}}\Big\|_{\max} > \max\Big\{t_1(T,d), t_2(T,d)\Big\}\Big) \leq 24\alpha^2,$$

when T and d are large enough. This completes the proof of (4.6). Combining the above inequality and Lemma 14 proves (4.8). □

Proof of Theorem 23. Theorem 23 follows immediately from Theorem 22 and Lemma 14. □

4.5 Discussion

In this section, we summarize the main contributions of this work, regarding the uniqueness as well as the generality of the Kolmogorov dependence condition.

The Kolmogorov dependence condition is closely related to the Doukhan's weak dependence conditions (Doukhan and Louhichi, 1999; Kallabis and Neumann, 2006; Doukhan and Neumann, 2007, 2008) developed for concentration inequalities. However, these conditions are not directly applicable for analyzing quantile-based robust statistics, since they are not invariant to non-smooth transformations of the stochastic process. In comparison, the Kolmogorov dependence condition developed in this paper is conveniently adapted to the non-smooth structure of quantile statistics. The Kolmogorov dependence condition also resembles the α-mixing conditions (Dedecker and Prieur, 2004; Kontorovich et al., 2008;

Merlevède et al., 2009, 2011) regarding the form of dependence measure. The key difference is that the dependence measure in α-mixing is defined in terms of σ-fields, which make the α-mixing conditions difficult to verify. In comparison, the Kolmogorov dependence condition relaxes the requirement for σ-fields, and is easily verified under many popular weak dependence conditions including the α-mixing conditions themselves.

The Kolmogorov dependence condition provides us a fairly general understanding of dependence. It serves as a necessary condition of a number of other weak dependence conditions, including VAR models, physical dependence conditions, mixing conditions, and various induced conditions from Doukhan's weak dependence condition. Thus, the theoretical results obtained under the Kolmogorov dependence condition shed light on the properties of other dependence conditions as well.

Chapter 5

Discussion and Future Work

High dimensional time series commonly arise in many scientific and economic areas. They present uniques challenges due to their high dimensionality, serial dependence, and many other domain-specific characteristics. In this research, we consider three specific settings of high dimensional time series: multiple time series with varying distributions, heavy-tailed time series, and a general time series with a novel dependence measure.

The first setting is motivated by the structure of the data from an fMRI study, where multiple subjects produce multiple time series with different covariance structures. We propose a kernel-based estimator for the graphical model of any subject, and derive theory on the consistency of the estimator. Our contributions lie in two aspects. First, our theory quantifies the strength one can borrow from across subjects in estimating the graphical model of any one subject. Secondly, we explicitly characterize the effect of the vector autoregressive structure by the ℓ_2 norm of the transition matrix. These results establish

a clear and rigorous understanding of the interactions between intra-subject information, inter-subject information, and serial dependence. On the other hand, from a methodological perspective, the assumed dependence structure itself is not exploited in the estimation procedure. It would an interest track of future work to explore how dependence structures can be used to improve estimation accuracy.

The second setting naturally arises in financial return data, where extreme events are common. We propose a novel formulation of portfolio optimization that accommodates arbitrarily heavy-tailed distributions for the returns of the candidate assets. The proposed method is innovative in that it establishes a novel risk metric of a portfolio, which naturally accommodates heavy-tailed distributions by using quantile statistics. The proposed method is also generic in that it is based on a generic scatter matrix that is not specific to any structures of the financial market. Alternative estimators that exploit the factor structures of financial asset prices have been proven successful. For future work, it's desirable to explore regularizations of the proposed scatter matrix according to these market-specific structures.

The third setting is motivated by the theoretical difficulty of analyzing quantile-based scatter matrix estimators using existing models of serial dependence. We propose a novel dependence condition called the Kolmogorov dependence, and showed that it naturally couples with the structure of quantile-based statistics. Moreover, the connections between Kolmogorov dependence and many other widely used dependence models are established and well understood. This not only makes our analysis of quantile-based statistics obtained

under Kolmogorov dependence fairly general, but also establishes a unified view over different concepts of serial dependence. For future work, it would be exciting to analyze the performance of many other statistical methods under Kolmogorov dependence, since the results would immediately shed light on the properties of other dependence models as well.

The three components of this thesis jointly provide a novel demonstration of the variation and integration of different dependence models in high dimensional data analysis. In particular, we develop different techniques for analyzing VAR models and mixing conditions that are highly specific to the statistical models and methodologies in question. On the other hand, these two conditions of serial dependence, along with many others, are unified under Kolmogorov dependence in that they are re-expressed in the common language of Kolmogorov dependence condition. The commonality sheds light on the fundamentals shared by all these difference dependence models.

Appendix A

Appendix to Chapter 2

A.1 Auto-Correlation and Cross-Correlation

In this section, we investigate the effect of the sign and strength of auto-correlation and cross-correlation on the rate of convergence. In detail, we define the diagonal entries of $\mathbf{A}(u)$ to be the auto-correlation coefficients, since they capture how $(\boldsymbol{x}_{it})_j$ depends on $\{\boldsymbol{x}_{i(t-1)}\}_j$, for $i = 1, \ldots, n$, $t = 2, \ldots, T$, and $j = 1, \ldots, d$. We define the off-diagonal entries of $\mathbf{A}(u)$ to be the cross-correlation coefficients, since they capture how $(\boldsymbol{x}_{it})_j$ depends on $\{\boldsymbol{x}_{i(t-1)}\}_{\setminus j}$. Since a general analysis is intractable, we focus on several special structures on $\mathbf{A}(u)$. We suppress the label u in $\mathbf{A}(u)$, and subject index i in $\boldsymbol{x}_{it}$ for notational brevity.

1. We first study the effect of auto-correlation. For highlighting autocorrelation alone, we set the cross-correlation coefficients to be 0 and consider the case where $\mathbf{A}$ is

diagonal: $\mathbf{A} = \mathrm{diag}(\rho_1, \ldots, \rho_d)$. This scenario is equivalent to d independent time series.

2. Secondly, we study the effect of the cross-correlation. To this end, we set the diagonal entries of $\mathbf{A}$ to be 0. In this scenario, at any time point, a variable does not depend on its value at the previous time point in the autoregression. Below we focus on two special structures on the off-diagonal entries, as exploited in Han and Liu (2013b).

   (a) $\mathbf{A}$ has a "band" structure, i.e., $\mathbf{A}_{ij} = \rho I(|i-j| = 1)$. In this case, the j-th entry of $\boldsymbol{x}_t$ only depends on adjacent entries at time $t-1$, i.e., entries in $\boldsymbol{x}_{t-1}$ with index differing from j by 1.

   (b) $\mathbf{A}$ is block diagonal. Each block has an "AR" structure. Specifically, let $\mathbf{A} = \mathrm{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_k)$, where $\mathbf{A}_l \in \mathbb{R}^{d_l \times d_l}$ for $l = 1, \ldots, k$. We have $(\mathbf{A}_l)_{ij} = \rho^{|i-j|} I(i \neq j)$, for $i, j = 1, \ldots, d_l$. In this case, the entries of $\boldsymbol{x}_t$ form k clusters. Temporal dependence occurs only within clusters. In each cluster, the cross-correlation coefficients decrease exponentially with the gap in index.

The next theorem summarizes the impact of the correlation coefficients on the rate of convergence.

Theorem A.1.1. *Let* $\mathbf{A}$ *be one of the transition matrices defined in* (1)*,* (2).i *and* (2).ii*. Inheriting the assumptions and notations in Lemma 1, we have:*

APPENDICES

(1). Under Scenario (1)*, we have*

$$\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left[\left\{\frac{\xi \sup_{u\in[0,1]} \|\mathbf{\Sigma}(u)\|_2}{1 - \max_{j=1,\ldots,d}(|\rho_j|)}\sqrt{\frac{\log d}{Tn}}\right\}^{1/2} + n^{-\frac{2}{2+\eta}}\right].$$

Thus, the magnitude of the maximum auto-correlation coefficient has a negative effect on the convergence rate. In comparison, the signs of the auto-correlation coefficients has no effect.

(2). Under Scenario (2). i*, we have*

$$\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left[\left\{\frac{\xi \sup_{u\in[0,1]} \|\mathbf{\Sigma}(u)\|_2}{1 - 2|\rho|\cos\{\pi/(d+1)\})}\sqrt{\frac{\log d}{Tn}}\right\}^{1/2} + n^{-\frac{2}{2+\eta}}\right].$$

Under Scenario (2).ii*, we have* $\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P[\alpha(\rho, \xi, \mathbf{\Sigma}, T, n, d)]$*, where* α*, as a function of* ρ*, is symmetric around* 0 *and monotonically increasing in for* $\rho > 0$*. Thus, the magnitude of the cross-correlation coefficients has a negative effect on the convergence rate. Again, the signs of the cross-correlation coefficients has no effect.*

Although Theorem A.1.1 only presents the effect of the correlation coefficients on the upper bound of estimation error, the simulation study in Section A.2.1 provides consistent results in estimation accuracy.

A.2 Additional Experiments

A.2.1 Impact of Temporal Dependence

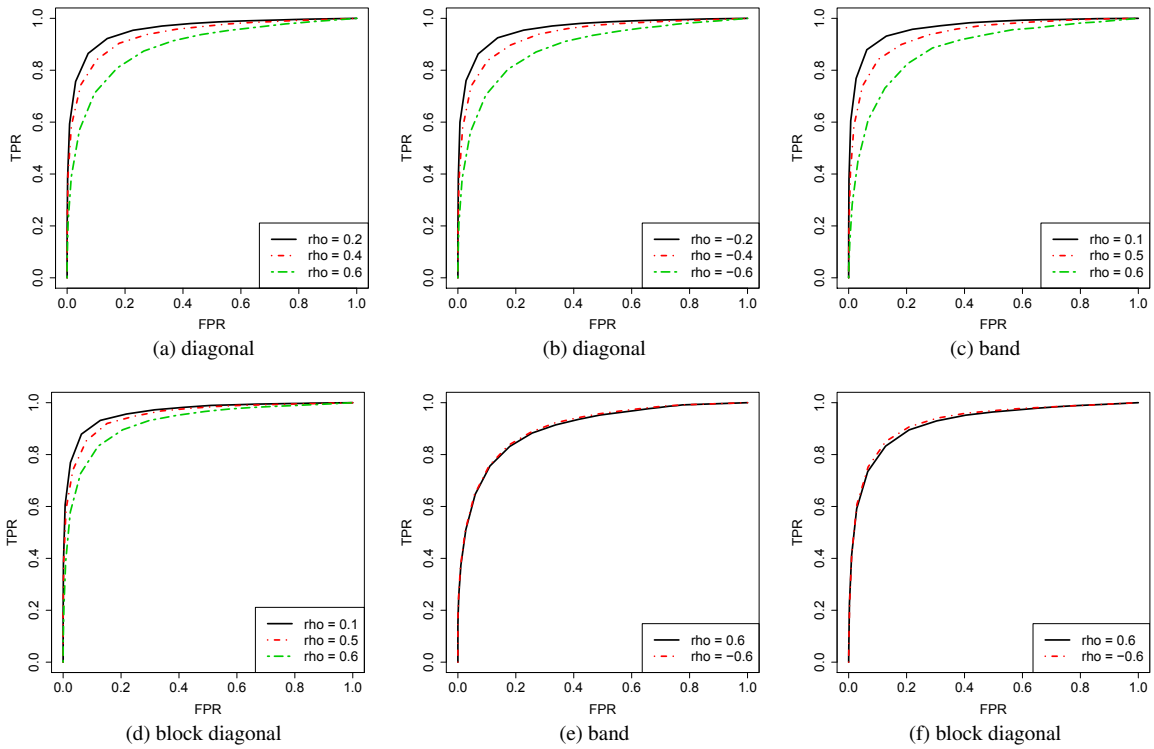
In this section, we investigate the impact of temporal dependence on graph estimation accuracy. Corresponding to the discussions in Section A.1, we consider three special structures of the transition matrix $\mathbf{A}(u) \in \mathbb{R}^{d\times d}$ to demonstrate the impact of auto-correlation and cross-correlation. To be illustrative, we fix the dimension $d = 10$. For simplicity, we let $\mathbf{A}(u)$ be constant over $u \in [0, 1]$, and suppress the label u in $\mathbf{A}(u)$.

1. diagonal: $\mathbf{A} = \mathrm{diag}(\rho, \ldots, \rho)$;

2. band: $\mathbf{A}_{ij} = \rho I(|i - j| = 1)$;

3. block diagonal: $\mathbf{A} = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3)$, where $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{3\times 3}$, and $\mathbf{A}_3 \in \mathbb{R}^{4\times 4}$, and $(\mathbf{A}_l)_{ij} = \rho^{|i-j|} I(i \neq j)$, for $l = 1, 2, 3$.

Using these transition matrices, we generated data according to Setting 1 described in Section 2.4.1.1. We fixed $n = 51$, $T = 50$, and $d = 10$, and target at label $u_0 = 0$. To investigate the impact of strong versus weak auto-correlation, we range ρ in $\{0.2, 0.4, 0.6\}$ and $\{-0.2, -0.4, -0.6\}$ under Scenario (1). Figures A.1(a) and A.1(b) display the results. One can see that large values of $|\rho|$ correspond to low estimation accuracy. Comparing Figures A.1(a) and A.1(b), it can be seen that the sign of the auto-correlation coefficients does not noticeably affect the ROC curves.



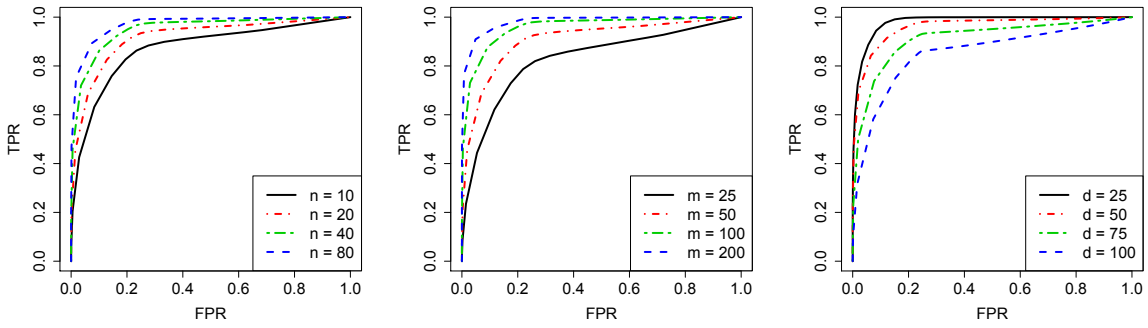
(a) diagonal (b) diagonal (c) band

(d) block diagonal (e) band (f) block diagonal

Figure A.1: ROC curves of KSE under three structured transition matrices: diagonal, band, and block diagonal. Data are synthesized under Setting 1. We set dimension $d = 10$; number of labels $n = 51$; number of observations $T = 50$.

To investigate the impact of strong versus weak positive cross-correlation, we vary ρ in $\{0.1, 0.5, 0.6\}$ under Scenarios (2) and (3). To keep $\|\mathbf{A}\|_2 < 1$, we scale $\mathbf{A}$ by $0.95/\|\mathbf{A}_{\max}\|_2$, where $\mathbf{A}_{\max}$ is the transition matrix when $\rho = 0.6$. Figures A.1(c) and A.1(d) show the results. Again, larger correlation results in decreased estimation accuracy.

Finally, to investigate the impact of strong positive versus strong negative cross-correlation, we compare $\rho = 0.6$ with $\rho = -0.6$ under Scenarios (2) and (3). Figures A.1(e) and A.1(f) deliver the results. Still the sign of cross-correlation does not dramatically affect the performance.



(a) n (b) T (c) d

Figure A.2: ROC curves of KSE under Setting 1 with varying label size n, sample size T, or dimension d.

A.2.2 Impact of Label Size n, Sample Size T, and Dimension d.

In this section, we empirically demonstrate how the label size n, sample size T, and dimension d may affect estimation accuracy. We inherit Setting 1 described in Section 2.4.1.1. We range n in $\{10, 20, 40, 80\}$, T in $\{25, 50, 100, 200\}$, and d in $\{25, 50, 75, 100\}$. Note that when d varies, n_{fix}, n_{grow}, and n_{decay} are scaled to maintain the same sparsity. Figure A.2 shows the results. As indicated by the rate of convergence in Section 2.3, estimation accuracy drops as we decrease n or T, or increase d.

The simulation results in Sections A.2.1 and A.2.2 provide empirical support for Theorem 1. Although only an upper bound on the estimation error is presented in Theorem 1, the rate of convergence does provide informative guidance on how the parameters may affect estimation accuracy.

A.2.3 Additional Results on ADHD-200 Data

A.2.3.1 Development of Brain Network Density

In this section, we investigate how brain network density changes with age. The number of edges in the estimated graph is controlled by λ. As Theorem 1 indicates, the proper choice of λ across the age spectrum depends on the heterogeneity of the multiple time series available. In detail, both the distribution of the subject ages and the number of observations under each subject affect the proper choice of λ. In order that the same λ is applicable across the age spectrum, we take a pre-processing step to achieve homogeneity.

To control the number of observations, T, we select the subjects with no fewer than 120 scans. We use only the first 120 scans of these subjects. To make sure that the subjects are distributed uniformly across the age spectrum, we subsampled 46 of the selected subjects whose ages form an equally spaced grid between 10 and 15. We abandon the ranges $[7.09, 10]$ and $[15, 21.83]$, since subjects are distributed rather heterogeneously across these ranges and do not fit into the grid.

Using the subsample of subjects, we can fix λ and estimate the brain networks at 26 target ages equally spaced across $[11, 14]$. We do not target at ages close to the boundaries, because fewer subjects are available around these boundaries. Figure A.3 demonstrates the estimated number of edges as a function of age, under three choices of λ. We observe that the estimated brain network density grows with age.

We note that although we removed possible confounding effects of sampling hetero-

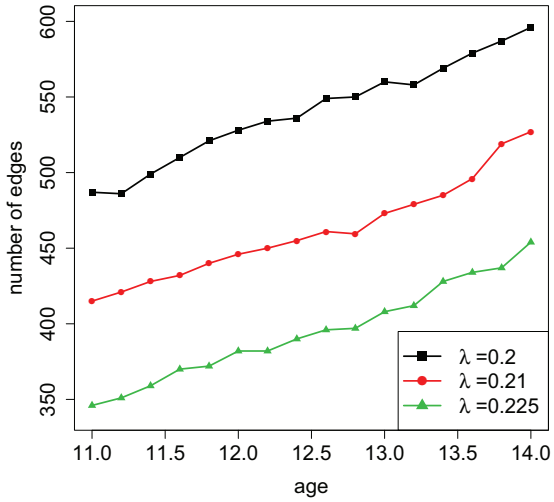


Figure A.3: The growth of estimated brain network density over age under three choices of λ. A subsample of the subjects from the ADHD-200 data are used to control λ.

geneity on the estimated network density, the proposed method still doesn't distinguish between the changes in brain complexity and the changes in structural heterogeneity over age. To address this issue, an assessment of confidence on the estimated numbers of edges across age is desired. That falls into the subject of statistical inference on high dimensional graphical models, which is an interesting area for future study.

A.2.3.2 The Impact of Bandwidth

In this section, the impact of bandwidths on estimation is considered. In practice, the bandwidth can be regarded as the degree of tradeoff between the label-specific networks and the population level networks. Under such a logic, a higher value of bandwidth will result in incorporating more information from the data points in other labels, and lead to an estimate closer to a population-level graph. This population-level graph will highlight the

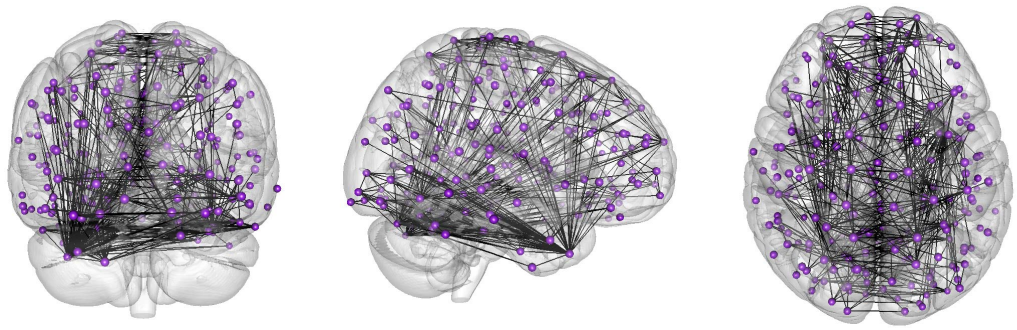
similarity between different graphs, while tending to ignore the label-specific differences. To illustrate this phenomenon empirically, consider estimating the brain network at age 21.83. We increase the bandwidth h, while setting all the other parameters fixed. As h is increased from 0.5 to 3, the weights in Equation (2.4) tends to be homogeneous across ages. Thus the graph ranges from age-specific level to the population level. Figure A.4 plots the different brain connectivity graphs estimated using different bandwidths.

There are two main discoveries: (i) The number of edges decreases to a population level of 674 as h increase to 3. This is intuitive, because the population level brain network will summarize the information across different levels and thus should be more concrete. (ii) When $h = 3$, the estimated brain network is close to the network estimated at age 7.09 shown in Figure 2.3 with most edges taking place at the occipital lobe region. This is expected because the occipital lobe region is the only part that has been well developed across the entire range of ages.

A.3 Techinical Proofs

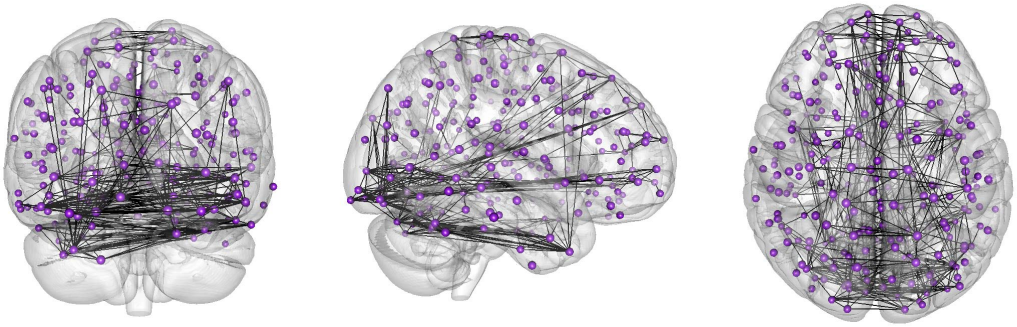
A.3.1 Proof of Lemma 1

The proof of Lemma 1 can be decomposed into two parts. In the first part, we prove that the bias term, $\mathbb{E}\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)$, can be controlled by the number of subjects n and bandwidth h. The result is provided in the following lemma.



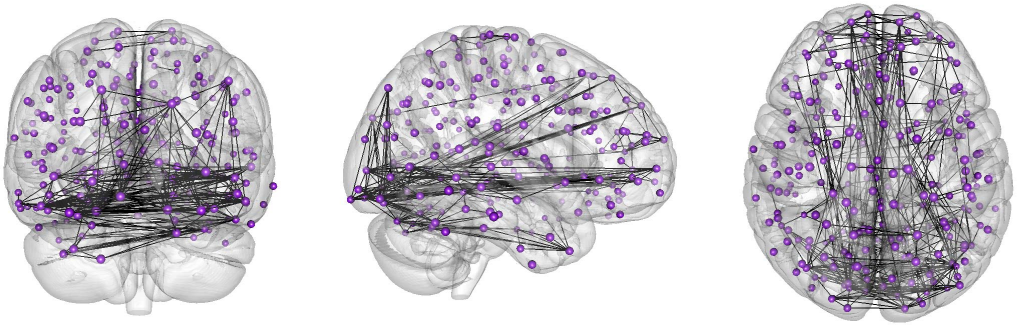
(a) coronal (b) sagittal (c) transverse

$h = 0.5$; 1032 edges



(d) coronal (e) sagittal (f) transverse

$h = 0.1$; 718 edges



(g) coronal (h) sagittal (i) transverse

$h = 3$; 674 edges

Figure A.4: Estimated brain connectivity network at age 21.83 among healthy subjects. The kernel bandwidth h takes the value 0.5, 1, 3, resulting to different brain connectivity networks from closer to the age-specific level, to closer to the population level.

Lemma A.3.1. *Supposing that the conditions in Lemma 1 hold, we have*

$$\max_{j,k}\left|\mathbb{E}\{\mathbf{S}(u_0)\}_{jk}-\Sigma_{jk}(u_0)\right|=O\left(h+\frac{1}{n^2h^{1+\eta}}\right).$$

Proof. By the definition of $\mathbf{S}(u_0)$ in Equation (2.3), we have

$$\mathbf{S}(u_0)=\sum_{i=1}^{n}\omega_i(u_0,h)\frac{1}{T}\sum_{k=1}^{T}\boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^{\mathsf{T}}.$$

Accordingly, we have

$$\begin{aligned}\mathbb{E}[\mathbf{S}(u_0)]_{jk}&=\sum_{i=1}^{n}\omega_i(u_0,h)\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^{\mathsf{T}}\\&=\sum_{i=1}^{n}\omega_i(u_0,h)\Sigma_{jk}(u_i)\\&=\frac{c(u_0)}{nh}\sum_{i=1}^{n}K\left(\frac{u_i-u_0}{h}\right)\Sigma_{jk}(u_i).\end{aligned}\tag{A.1}$$

By Theorem 1.1 in Tasaki (2009) and Assumption **(A2)**, we have

$$\begin{aligned}&\frac{c(u_0)}{nh}\sum_{i=1}^{n}K\left(\frac{u_i-u_0}{h}\right)\Sigma_{jk}(u_i)\\=&\frac{c(u_0)}{h}\int_0^1 K\left(\frac{u-u_0}{h}\right)\Sigma_{jk}(u)du+O\left[\frac{c(u_0)}{n^2h}\sup_{u\in[0,1]}\frac{d^2}{du^2}\left\{K\left(\frac{u-u_0}{h}\right)\Sigma_{jk}(u)\right\}\right]\\=&c(u_0)\int_{-\frac{u_0}{h}}^{\frac{1-u_0}{h}}K(u)\Sigma_{jk}(u_0+hu)du+O\left(\frac{1}{n^2h^{1+\eta}}\right)\\=&c(u_0)\int_{a(u_0)}^{b(u_0)}K(u)\left\{\Sigma_{jk}(u_0)+hu\Sigma'_{jk}(\zeta)\right\}du+O\left(\frac{1}{n^2h^{1+\eta}}\right),\end{aligned}\tag{A.2}$$

where $a(u_0) := -I(u_0 \in (0,1])$, $b(u_0) := I(u_0 \in [0,1))$, $\Sigma'_{jk}(u) := \frac{d}{du}\Sigma_{jk}(u)$, and ζ lies between u_0 and $u_0 + hu$. The last equality is because $h \to 0$ and $K(u)$ has support $[-1,1]$. By Equation (2.2), we have

$$c(u_0)\int_{a(u_0)}^{b(u_0)} K(u)\Sigma_{jk}(u_0)du = \Sigma_{jk}(u_0). \tag{A.3}$$

By Equation (2.2) and Assumption **(A1)**, we have

$$\begin{aligned}&\left|c(u_0)\int_{a(u_0)}^{b(u_0)} K(u)hu\Sigma'_{jk}(\zeta)du\right| \le C_2 h\left|c(u_0)\int_{a(u_0)}^{b(u_0)} |u|K(u)du\right| \\ &= 2C_2 h\left|\int_0^1 uK(u)du\right| = O(h).\end{aligned} \tag{A.4}$$

Combining (A.1), (A.2), (A.3), and (A.4), we have

$$\left|\mathbb{E}\{\mathbf{S}(u_0)\}_{jk} - \Sigma_{jk}(u_0)\right| = O\left(h + \frac{1}{n^2h^{1+\eta}}\right).$$

This completes the proof. □

We then proceed to the second lemma, which provides an upper bound of the distance between the estimator $\mathbf{S}(u_0)$ and its expectation $\mathbb{E}\mathbf{S}(u_0)$.

Lemma A.3.2. *Supposing that the conditions in Lemma 1 hold, we have*

$$\max_{j,k}\left|\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}\right| = O_P\left[\frac{\xi \cdot \sup_{u\in[0,1]}\|\mathbf{\Sigma}(u)\|_2}{h\{1-\sup_{u\in[0,1]}\|\mathbf{A}(u)\|_2\}}\sqrt{\frac{\log d}{Tn}}\right].$$

Proof. For $i = 1, \ldots, n$ and $t = 1, \ldots, T$, let $\boldsymbol{y}_{it} := (y_{it1}, \ldots, y_{itd})^{\mathsf{T}}$ be a d-dimensional random vector with $y_{itj} = x_{itj}/\sqrt{\Sigma_{jj}(u_i)}$. Define correlation coefficient $\rho_{jk}(u_i) := \Sigma_{jk}(u_i)/\sqrt{\Sigma_{jj}(u_i)\Sigma_{kk}(u_i)}$. We then have

$$
\begin{aligned}
&\mathbb{P}\left[|\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}| > \epsilon\right] \\
=&\mathbb{P}\left[\left|\sum_{i=1}^{n} \omega_i(u_0, h)\left\{\frac{1}{T}\sum_{t=1}^{T} x_{itj}x_{itk} - \Sigma_{jk}(u_i)\right\}\right| > \epsilon\right] \\
=&\mathbb{P}\left\{\left|\sum_{i=1}^{n} \omega_i(u_0, h)\sqrt{\Sigma_{jj}(u_i)\Sigma_{kk}(u_i)}\left(\left[\frac{1}{T}\sum_{t=1}^{T}(y_{itj} + y_{itk})^2 - 2\{1 + \rho_{jk}(u_i)\}\right]\right.\right. \\
&\left.\left. - \left[\frac{1}{T}\sum_{t=1}^{T}(y_{itj} - y_{itk})^2 - 2\{1 - \rho_{jk}(u_i)\}\right]\right)\right| > 4\epsilon\right\} \\
\leq&\mathbb{P}\left\{\left|\sum_{i=1}^{n} \omega_i^*(u_0, h)\left(\left[\frac{1}{T}\sum_{t=1}^{T}(y_{itj} + y_{itk})^2 - 2\{1 + \rho_{jk}(u_i)\}\right]\right)\right| > 2\epsilon\right\} \\
&+ \mathbb{P}\left\{\left|\sum_{i=1}^{n} \omega_i^*(u_0, h)\left(\left[\frac{1}{T}\sum_{t=1}^{T}(y_{itj} - y_{itk})^2 - 2\{1 - \rho_{jk}(u_i)\}\right]\right)\right| > 2\epsilon\right\} \\
:=&P_1 + P_2, \qquad \text{(A.5)}
\end{aligned}
$$

where $\omega_i^*(u_0, h) := \omega_i(u_0, h)\sqrt{\Sigma_{jj}(u_i)\Sigma_{kk}(u_i)}$.

Let $\boldsymbol{Z} := (\boldsymbol{Z}_1^{\mathsf{T}}, \ldots, \boldsymbol{Z}_n^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{nT}$, where $\boldsymbol{Z}_i := (y_{i1j}+y_{i1k}, y_{i2j}+y_{i2k}, \ldots, y_{iTj}+y_{iTk})^{\mathsf{T}}$.

We have $\boldsymbol{Z}_{i_1}$ is independent of $\boldsymbol{Z}_{i_2}$ for any $i_1 \neq i_2$. Let

$$\mathbf{B} := \begin{pmatrix} \sqrt{\omega_1^*(u_0,h)} \cdot \mathbf{I}_T & 0 & \dots & 0 \\ 0 & \sqrt{\omega_2^*(u_0,h)} \cdot \mathbf{I}_T & & 0 \\ & & \ddots & \\ 0 & 0 & & \sqrt{\omega_n^*(u_0,h)} \cdot \mathbf{I}_T \end{pmatrix}$$

be a Tn by Tn diagonal matrix. Then we can rewrite P_1 as $P_1 = \mathbb{P}(|\ \|\mathbf{B}\boldsymbol{Z}\|_2^2 - E\|\mathbf{B}\boldsymbol{Z}\|_2^2\ | > 2T\epsilon)$. Using the property of Gaussian distribution, we have $\mathbf{B}\boldsymbol{Z} \sim N_{Tn}(\mathbf{0}, \mathbf{Q})$, where $\mathbf{Q} := \mathbf{B}\mathrm{cov}(\boldsymbol{Z})\mathbf{B}$ and

$$\mathrm{cov}(\boldsymbol{Z}) = \begin{pmatrix} \mathrm{cov}(\boldsymbol{Z}_1) & 0 & \dots & 0 \\ 0 & \mathrm{cov}(\boldsymbol{Z}_2) & & 0 \\ & & \ddots & \\ 0 & 0 & & \mathrm{cov}(\boldsymbol{Z}_n) \end{pmatrix}.$$

Let $\{\mathrm{cov}(\boldsymbol{Z}_i)\}_{pq}$ be the (p,q) element of $\mathrm{cov}(\boldsymbol{Z}_i)$. We have

$$\begin{aligned} |\{\mathrm{cov}(\boldsymbol{Z}_i)\}_{pq}| &= |\mathrm{cov}(y_{ipj} + y_{ipk}, y_{iqj} + y_{iqk})| \\ &= |\mathrm{cov}(y_{ipj}, y_{iqj}) + \mathrm{cov}(y_{ipj}, y_{iqk}) + \mathrm{cov}(y_{ipk}, y_{iqj}) + \mathrm{cov}(y_{ipk}, y_{iqk})| \\ &\leq \frac{|\mathrm{cov}(x_{ipj}, x_{iqj}) + \mathrm{cov}(x_{ipj}, x_{iqk}) + \mathrm{cov}(x_{ipk}, x_{iqj}) + \mathrm{cov}(x_{ipk}, x_{iqk})|}{\min_r \Sigma_{rr}(u_i)} \\ &\leq \frac{4\|\mathbf{A}(u_i)\|_2^{|p-q|}\|\mathbf{\Sigma}(u_i)\|_2}{\min_r \Sigma_{rr}(u_i)}. \end{aligned}$$

The last inequality is due to the property of the VAR(1) models. Thus

$$\begin{aligned}\|\mathbf{Q}\|_2 \leq & \max_{1\leq s\leq Tn}\sum_{r=1}^{Tn}|\mathbf{Q}_{sr}| \\ = & \max_{i=1,\ldots,n;p=1,\ldots,T}\sum_{q=1}^{T}\omega_i^*(u_0,h)|\{\text{cov}(\boldsymbol{Z}_i)\}_{pq}| \\ \leq & \max_{i=1,\ldots,n}\omega_i^*(u_0,h)\frac{4\|\boldsymbol{\Sigma}(u_i)\|_2}{\min_r\Sigma_{rr}(u_i)}\cdot 2\sum_{q=0}^{\infty}\|\mathbf{A}(u_i)\|_2^q \\ \leq & \frac{16C_1}{nh}\cdot\frac{\xi\sup_{u\in[0,1]}\|\boldsymbol{\Sigma}(u)\|_2}{1-\sup_{u\in[0,1]}\|\mathbf{A}(u)\|_2}. \end{aligned} \tag{A.6}$$

The last inequality is due to the fact that $\omega_i^*(u_0,h) = \omega_i(u_0,h)\sqrt{\Sigma_{jj}(u_i)\Sigma_{kk}(u_i)} \leq \frac{2}{nh}\cdot \sup_v K(v)\cdot\sup_u\max_r\Sigma_{rr}(u)$.

Finally, using Lemma I.2 in Negahban and Wainwright (2011), we have

$$\begin{aligned}\mathbb{P}(|\ \|\mathbf{B}\boldsymbol{Z}\|_2^2-\mathbb{E}\|\mathbf{B}\boldsymbol{Z}\|_2^2\ |>2T\epsilon) \leq & 2\exp\left\{-\frac{Tn}{2}\left(\frac{\epsilon}{2n\|\mathbf{Q}\|_2}-\frac{2}{\sqrt{Tn}}\right)^2\right\}+2\exp\left(-\frac{Tn}{2}\right) \\ \leq & 4\exp\left\{-\frac{Tn}{2}\left(\frac{\epsilon}{4n\|\mathbf{Q}\|_2}\right)^2\right\}, \end{aligned} \tag{A.7}$$

for large enough n.

Using the same technique, we can show that P_2 in Equation (A.5) can also be controlled by the bound in (A.7). So using the union bound, we have

$$\begin{aligned}\mathbb{P}\left[\max_{j,k}|\{\mathbf{S}(u_0)\}_{jk}-\mathbb{E}\{\mathbf{S}(u_0)\}_{jk}|>\epsilon\right] \leq & \sum_{j,k}\mathbb{P}\left[|\{\mathbf{S}(u_0)\}_{jk}-\mathbb{E}\{\mathbf{S}(u_0)\}_{jk}|>\epsilon\right] \\ \leq & 8d^2\exp\left(-\frac{T\epsilon^2}{32n\|\mathbf{Q}\|_2^2}\right). \end{aligned} \tag{A.8}$$

Thus, using Equations (A.6) and (A.8), we have

$$\begin{aligned}\max_{j,k} |\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}| &= O_P\left(\|\mathbf{Q}\|_2\sqrt{\frac{n\log d}{T}}\right) \\ &= O_P\left[\frac{\xi\cdot\sup_{u\in[0,1]}\|\mathbf{\Sigma}(u)\|_2}{h\{1-\sup_{u\in[0,1]}\|\mathbf{A}(u)\|_2\}}\sqrt{\frac{\log d}{Tn}}\right].\end{aligned}$$

This completes the proof. □

A.3.1.1 Proof of Lemma 1

The rate of convergence in Lemma 1 can be obtained by balancing the convergence rates in Lemmas A.3.1 and A.3.2. More specifically, we first have

$$\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} \leq \|\mathbf{S}(u_0) - \mathbb{E}\mathbf{S}(u_0)\|_{\max} + \|\mathbb{E}\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max}.$$

For notational brevity, we denote $\theta := \xi\sup_{u\in[0,1]}\|\mathbf{\Sigma}(u)\|_2/\{1-\sup_{u\in[0,1]}\|\mathbf{A}(u)\|_2\}$. We then have

$$\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left(h + \frac{1}{n^2h^{1+\eta}} + \frac{\theta}{h}\sqrt{\frac{\log d}{Tn}}\right).$$

We first balance the first and third terms in the above upper bound, having that

$$h = \frac{\theta}{h}\sqrt{\frac{\log d}{Tn}} \Rightarrow h = \left(\theta\sqrt{\frac{\log d}{Tn}}\right)^{1/2}.$$

We then balance the first and second terms, and have that

$$h = \frac{1}{n^2 h^{1+\eta}} \Rightarrow h = n^{-\frac{2}{2+\eta}}.$$

Based on the above two results, we have that, on one hand, if $\left(\theta\sqrt{\frac{\log d}{Tn}}\right)^{1/2} > n^{-\frac{2}{2+\eta}}$, we can set

$$h = \left(\theta\sqrt{\frac{\log d}{Tn}}\right)^{1/2}.$$

Then we have

$$h = \frac{\theta}{h}\sqrt{\frac{\log d}{Tn}} > \frac{1}{n^2 h^{1+\eta}} \Rightarrow \|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left\{\left(\theta\sqrt{\frac{\log d}{Tn}}\right)^{1/2}\right\}. \quad \text{(A.9)}$$

On the other hand, if $\left(\theta\sqrt{\frac{\log d}{Tn}}\right)^{1/2} \leq n^{-\frac{2}{2+\eta}}$, we can set

$$h = n^{-\frac{2}{2+\eta}}.$$

Then we have

$$h = \frac{1}{n^2 h^{1+\eta}} \geq \frac{\theta}{h}\sqrt{\frac{\log d}{Tn}} \Rightarrow \|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left(n^{-\frac{2}{2+\eta}}\right). \quad \text{(A.10)}$$

Combining (A.9) and (A.10), we have the desired result.

A.3.2 Proof of Theorem A.1.1

The following two lemmas are needed in the proof of Theorem A.1.1.

Lemma A.3.3. *Let* $\mathbf{M}_\rho \in \mathbb{R}^{d\times d}$ *be a matrix where* $\mathbf{M}_{jk} = \rho^{|j-k|} I(j \neq k)$. *Then* $\mathbf{M}_\rho$ *and* $\mathbf{M}_{-\rho}$ *have the same set of eigenvalues.*

Proof. Let $\mathbf{B} \in \mathbb{R}^{d\times d}$ be a diagonal matrix with $\mathbf{B}_{ii} = (-1)^i$. Noting that $(-1)^{i+j} = (-1)^{|i-j|}$ for all $i, j \in \{1, \ldots, d\}$, we have $\mathbf{M}_{-\rho} = \mathbf{B}\mathbf{M}\mathbf{B}^{-1}$. Thus $\mathbf{M}_\rho$ has the same set of eigenvalues as $\mathbf{M}_{-\rho}$. □

Lemma A.3.4. *Let* $\mathbf{N}_\rho \in \mathbb{R}^{d\times d}$ *be a matrix where* $\mathbf{N}_{jk} = \rho^{|j-k|}$ *and* $0 \leq \rho_1 \leq \rho_2$, *we have* $\|\mathbf{N}_{\rho_1}\|_2 \leq \|\mathbf{N}_{\rho_2}\|_2$.

Proof. $\mathbf{N}_{\rho_1}$ is the Hadamard product of $\mathbf{N}_{\rho_1/\rho_2}$ and $\mathbf{N}_{\rho_2}$:

$$\mathbf{N}_{\rho_1} = \mathbf{N}_{\rho_1/\rho_2} \circ \mathbf{N}_{\rho_2}.$$

By Theorem 5.3.4 of Roger and Charles (1994), any eigenvalue $\lambda(\mathbf{N}_{\rho_1/\rho_2} \circ \mathbf{N}_{\rho_2})$ of $\mathbf{N}_{\rho_1/\rho_2} \circ \mathbf{N}_{\rho_2}$ satisfies

$$\lambda(\mathbf{N}_{\rho_1/\rho_2} \circ \mathbf{N}_{\rho_2}) \leq (\max_{1\leq i\leq d} \mathbf{N}_{\rho_1/\rho_2})_{ii} \lambda_{\max}(\mathbf{N}_{\rho_2}) = \|\mathbf{N}_{\rho_2}\|_2.$$

Thus $\|\mathbf{N}_{\rho_1}\|_2 \leq \|\mathbf{N}_{\rho_2}\|_2$. □

A.3.2.1 Proof of Theorem A.1.1

Under Scenario (1), it is straightforward to have $\|\mathbf{A}\|_2 = \max_{j=1,\ldots,d} |\rho_j|$. Plugging it into Equation 2.9 proves the first part.

Under Scenario (2).i, it is well known that $\|\mathbf{A}\|_2 = 2|\rho| \cos\{\pi/(d+1)\}$. See, for example, Smith (1978) for details. This proves the second part.

Under Scenario (2).ii, the eigenvalues of $\mathbf{A}$ consist of the eigenvalues of each block. From Lemma A.3.3, we conclude that $\|\mathbf{A}\|_2$ do not depend on the sign of ρ. To prove monotonicity, note that $\|\mathbf{A}\|_2 = \max_{l=1,\ldots,k} \|\mathbf{A}_l\|_2$ and $\|\mathbf{A}_l\|_2 = \|\mathbf{N}_\rho - I_{d_l}\|_2 = \|\mathbf{N}_\rho\|_2 - 1$ for $\mathbf{N}_\rho \in \mathbb{R}^{d_l \times d_l}$. The desired result follows from Lemma A.3.4.

A.3.3 Proof of Lemma 2

To prove Lemma 2, we need an improved upper bound on the distance between $\mathbf{S}(u_0)$ and $\mathbb{E}\mathbf{S}(u_0)$. We provide such a result in Lemma A.3.5. The proof of Lemma A.3.5 can be regarded as an extension to the proof of Lemma 6 in Zhou et al. (2010).

Lemma A.3.5. *Suppose that Assumptions* **(B1), (B2)**, *and* **(B3)** *in Lemma 2 hold, and* $n^{-2/5} < h < 1$. *Then we have there exist absolute positive constants* C_4 *and* C_5, *such that for*

$$\epsilon < \frac{C_4\{\Sigma_{jj}^2(u_0)\Sigma_{kk}^2(u_0) + \Sigma_{jk}^2(u_0)\}}{\max_{i=1,\ldots,n} K\{(u_i - u_0)/h\}\Sigma_{jj}(u_i)\Sigma_{kk}(u_i)},$$

we have

$$\mathbb{P}\left[|\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}| > \epsilon\right] \leq 2\exp(-C_5 Tnh\epsilon^2).$$

Proof. By the definition of $\mathbf{S}(u_0)$, we have

$$\begin{aligned}
\mathbb{P}\left[|\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}| > \epsilon\right] =& \mathbb{P}\left[\left|\sum_{i=1}^{n} w_i(u_0,h)\left\{\frac{1}{T}\sum_{t=1}^{T} x_{itj}x_{itk} - \Sigma_{jk}(u_i)\right\}\right| > \epsilon\right] \\
\leq& \mathbb{P}\left[\sum_{i=1}^{n} w_i(u_0,h)\left\{\frac{1}{T}\sum_{t=1}^{T} x_{itj}x_{itk} - \Sigma_{jk}(u_i)\right\} > \epsilon\right] \\
&+ \mathbb{P}\left[\sum_{i=1}^{n} w_i(u_0,h)\left\{-\frac{1}{T}\sum_{t=1}^{T} x_{itj}x_{itk} + \Sigma_{jk}(u_i)\right\} > \epsilon\right] \\
:=& P_3 + P_4.
\end{aligned}$$

By Markov's inequality, $\forall r > 0$,

$$\begin{aligned}
P_3 &= \mathbb{P}\left(\exp\left[Tnr\sum_{i=1}^{n} w_i(u_0,h)\left\{\frac{1}{T}\sum_{t=1}^{T} x_{itj}x_{itk} - \Sigma_{jk}(u_i)\right\}\right] > e^{Tnr\epsilon}\right) \\
&\leq \frac{1}{e^{Tnr\epsilon}}\mathbb{E}\exp\left[r\sum_{i=1}^{n}\frac{2}{h}K\left(\frac{u_i-u_0}{h}\right)\sum_{t=1}^{T}\{x_{itj}x_{itk} - \Sigma_{jk}(u_i)\}\right] \\
&= e^{-Tnr\epsilon}\prod_{i=1}^{n}\exp\left\{-Tr\frac{2}{h}K\left(\frac{u_i-u_0}{h}\right)\Sigma_{jk}(u_i)\right\}\prod_{i=1}^{n}\left[\mathbb{E}\exp\left\{r\frac{2}{h}K\left(\frac{u_i-u_0}{h}\right)x_{itj}x_{itk}\right\}\right]^T.
\end{aligned}$$

The last equality is due to that $\{\boldsymbol{X}^{u_i}\}_{i=1}^n$ are independent and $\{\boldsymbol{x}_{it}\}_{t=1}^T$ are i.i.d.. Using the same technique, we can get similar result for P_4. The rest of the proof can be derived by following Lemma 6 in Zhou et al. (2010), where we replace n with Tn. Here the assump-

tion that $n^{-2/5} < h < 1$ and Assumption **(B2)** are required in the proof of Proposition 5 in Zhou et al. (2010). □

Using Lemma A.3.5, we can now proceed to prove Lemma 2. Because if the kernel function satisfies Assumption **(A2)** for some $\eta = \eta_1 > 0$, then this kernel function also satisfies Assumption **(A2)** for $\eta = \max(3, \eta_1)$, so without loss of generality, in the sequel we assume that $\eta \geq 3$ in Assumption **(A2)**.

A.3.3.1 Proof of Lemma 2

Using Lemma A.3.5, we have

$$\begin{aligned}\mathbb{P}\left[\max_{jk}|\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}| > \epsilon\right] &\leq \sum_{jk}\mathbb{P}\left[|\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}| > \epsilon\right] \\ &\leq \exp\left(2\log d - C_5 Tnh\epsilon^2\right),\end{aligned}$$

for $n^{-2/5} < h < 1$. Now setting $\epsilon = \sqrt{3\log d/(C_5 Tnh)}$, we have

$$\mathbb{P}\left[\max_{jk}|\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}| > \sqrt{\frac{3\log d}{C_5 Tnh}}\right] \leq \frac{1}{d}.$$

Accordingly, as $d \to \infty$, we have

$$\max_{jk}|\{\mathbf{S}(u_0)\}_{jk} - \mathbb{E}\{\mathbf{S}(u_0)\}_{jk}| = O_P\left(\sqrt{\frac{\log d}{Tnh}}\right).$$

Together with Lemma A.3.1, we have

$$\|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left(h + \frac{1}{n^2h^{1+\eta}} + \sqrt{\frac{\log d}{Tnh}}\right).$$

Similarly as the proof of Lemma 1, to balance the first and third terms, we set

$$h = \sqrt{\frac{\log d}{Tnh}} \Rightarrow h = \left(\frac{\log d}{Tn}\right)^{1/3}.$$

To balance the first and second terms, we set

$$h = \frac{1}{n^2h^{1+\eta}} \Rightarrow h = \frac{1}{n^{2/(2+\eta)}}.$$

If $\left(\frac{\log d}{Tn}\right)^{1/3} > \frac{1}{n^{2/(2+\eta)}}$, we set $h = \left(\frac{\log d}{Tn}\right)^{1/3}$. Then we have

$$h = \sqrt{\frac{\log d}{Tnh}} > \frac{1}{n^2h^{1+\eta}} \Rightarrow \|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left\{\left(\frac{\log d}{Tn}\right)^{1/3}\right\}. \tag{A.11}$$

Note that $\eta \geq 3$ implies that $h > n^{-2/(2+\eta)} > n^{-2/5}$.

If $\left(\frac{\log d}{Tn}\right)^{1/3} \leq \frac{1}{n^{2/(2+\eta)}}$, we set $h = \frac{1}{n^{2/(2+\eta)}}$. Then we have

$$h = \frac{1}{n^2h^{1+\eta}} \geq \sqrt{\frac{\log d}{Tnh}} \Rightarrow \|\mathbf{S}(u_0) - \mathbf{\Sigma}(u_0)\|_{\max} = O_P\left\{\frac{1}{n^{2/(2+\eta)}}\right\}. \tag{A.12}$$

Combining (A.11) and (A.12) we have the desired result.

Appendix B

Appendix to Chapter 3

B.1 Supporting Lemmas

We first derive the concentration inequality for the robust scale estimator $\hat{\sigma}^{\mathrm{Q}}$. It intrinsically relies on the concentration of the U-statistic,

$$U_T(\psi_u) := \frac{2}{T(T-1)} \sum_{1 \le s < t \le T} \psi_u(X_s, X_t), \tag{B.1}$$

for kernel function $\psi_u(x, y) := I(|x - y| \le u)$ under a ϕ-mixing process $\{X_t\}_{t \in \mathbb{Z}}$. To this end, we first focus on the bias and variance of $U_T(\psi_u)$.

Lemma 4. *Let $\{X_t\}_{t \in \mathbb{Z}}$ be a stationary ϕ-mixing process such that $\phi(n) \le n^{-1-\epsilon}$ for any $n > 0$ and some constant $\epsilon > 0$, and $\tilde{X}$ be an independent copy of X_1. Suppose X_1 is absolutely continuous. Denote by $G(u) := \mathbb{P}(|X_1 - \tilde{X}| \le u)$ the distribution function of*

$|X_1 - \tilde{X}|$. *For* $U_T(\psi_u)$ *defined in* (B.1)*, we have*

$$\left|\mathbb{E}U_T(\psi_u) - G(u)\right| \leq \frac{2C_\epsilon}{T},$$

for any $u > 0$*, where* $C_\epsilon = \sum_{k=1}^{\infty} 1/k^{1+\epsilon}$ *is a constant only depending on* ϵ*.*

Proof. Denote $G_{st}(u) := \mathbb{P}(|X_s - X_t| \leq u)$ to be the distribution function of $|X_s - X_t|$ for $s < t$. Let $M > 0$ be a constant and

$$-M = a_{-h}^{(h)} < \cdots < a_0^{(h)} < \cdots < a_h^{(h)} = M$$

be a sequence of real numbers satisfying

$$\max_{-h<k\leq h}(a_k^{(h)} - a_{k-1}^{(h)}) \leq u \text{ and } \lim_{h\to\infty} \max_{-h<k\leq h}(a_k^{(h)} - a_{k-1}^{(h)}) = 0. \tag{B.2}$$

Given $X_s \in [a_{k-1}^{(h)}, a_k^{(h)}]$, we have that $|X_s - X_t| \leq u$ implies $X_t \in [a_{k-1}^h - u, a_k^{(h)} + u]$. Thus, we have

$$\begin{aligned}
&\mathbb{P}(|X_s - X_t| \leq u, X_s \in [-M, M]) \\
&= \sum_{-h<k\leq h} \mathbb{P}(|X_s - X_t| \leq u \mid X_s \in [a_{k-1}, a_k])\mathbb{P}(X_s \in [a_{k-1}, a_k]) \\
&\leq \sum_{-h<k\leq h} \mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u] \mid X_s \in [a_{k-1}, a_k])\mathbb{P}(X_s \in [a_{k-1}, a_k]). \qquad \text{(B.3)}
\end{aligned}$$

On the other hand, given $X_s \in [a_{k-1}^{(h)}, a_k^{(h)}]$, we have $X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u]$ implies

$|X_s - X_t| \le u$. Thus, we have

$$\begin{aligned}
&\mathbb{P}(|X_s - X_t| \le u, X_s \in [-M, M]) \\
&= \sum_{-h<k\le h} \mathbb{P}(|X_s - X_t| \le u \mid X_s \in [a_{k-1}, a_k])\mathbb{P}(X_s \in [a_{k-1}, a_k]) \\
&\ge \sum_{-h<k\le h} \mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u] \mid X_s \in [a_{k-1}, a_k])\mathbb{P}(X_s \in [a_{k-1}, a_k]). \qquad \text{(B.4)}
\end{aligned}$$

Now define $\psi_h^U := \sum_{-h<k\le h} \mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u])\mathbb{P}(X_s \in [a_{k-1}, a_k])$, $\psi_h^L := \sum_{-h<k\le h} \mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u])\mathbb{P}(X_s \in [a_{k-1}, a_k])$, and

$$\psi_h := \begin{cases} \psi_h^L, & \text{if } \mathbb{P}(|X_s - X_t| \le u, X_s \in [-M, M]) > \psi_h^L; \\ \psi_h^U, & \text{otherwise.} \end{cases}$$

Note that $\psi_h^L \le \psi_h^U$. If $\mathbb{P}(|X_s - X_t| \le u, X_s \in [-M, M]) > \psi_h^L$, by the definition of ψ_h and (B.3), we have

$$\begin{aligned}
&|\mathbb{P}(|X_s - X_t| \le u, X_s \in [-M, M]) - \psi_h| = \mathbb{P}(|X_s - X_t| \le u, X_s \in [-M, M]) - \psi_h^L \\
&\le \sum_{-h<k\le h} \left|\mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u] \mid X_s \in [a_{k-1}, a_k]) - \mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u])\right| \\
&\qquad \mathbb{P}(X_s \in [a_{k-1}, a_k]) \\
&\le \sum_{-h<k\le h} \left|\mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u] \mid X_s \in [a_{k-1}, a_k]) - \mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u])\right| \\
&\qquad \mathbb{P}(X_s \in [a_{k-1}, a_k]) + \sum_{-h<k\le h} \left|\mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u]) - \mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u])\right|
\end{aligned}$$

$$\mathbb{P}(X_s \in [a_{k-1}, a_k])$$

$$\leq \phi(t-s) + \max_{-h<k\leq h} \left|\mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u]) - \mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u])\right|. \tag{B.5}$$

On the other hand, if $\mathbb{P}(|X_s - X_t| \leq u, X_s \in [-M, M]) \leq \psi_h^L$, since $\psi_h^L \leq \psi_h^U$, by the definition of ψ_h and (B.4), we have

$$|\mathbb{P}(|X_s - X_t| \leq u, X_s \in [-M, M]) - \psi_h| = \psi_h^U - \mathbb{P}(|X_s - X_t| \leq u, X_s \in [-M, M])$$

$$\leq \sum_{-h<k\leq h} \left|\mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u]) - \mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u] \mid X_s \in [a_{k-1}, a_k])\right|$$

$$\mathbb{P}(X_s \in [a_{k-1}, a_k])$$

$$\leq \sum_{-h<k\leq h} \left|\mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u] \mid X_s \in [a_{k-1}, a_k]) - \mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u])\right|$$

$$\mathbb{P}(X_s \in [a_{k-1}, a_k]) + \sum_{-h<k\leq h} \left|\mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u]) - \mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u])\right|$$

$$\mathbb{P}(X_s \in [a_{k-1}, a_k])$$

$$\leq \phi(t-s) + \max_{-h<k\leq h} \left|\mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u]) - \mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u])\right|. \tag{B.6}$$

Thus, combining (B.5) and (B.6), we have

$$|\mathbb{P}(|X_s - X_t| \leq u, X_s \in [-M, M]) - \psi_h|$$

$$\leq \phi(t-s) + \max_{-h<k\leq h} \left|\mathbb{P}(X_t \in [a_k^{(h)} - u, a_{k-1}^{(h)} + u]) - \mathbb{P}(X_t \in [a_{k-1}^{(h)} - u, a_k^{(h)} + u])\right|.$$

Let $h \to \infty$. Using (B.2) and the assumption that X_t is absolutely continuous, we have

$$\left|\mathbb{P}(|X_s - X_t| \leq u, X_s \in [-M, M]) - \int_{-M}^{M} \mathbb{P}(X_s \in [a-u, a+u]) d\mathbb{P}(X_s = a)\right| \leq \phi(t-s).$$

Now, let $M \to \infty$, we further obtain

$$\left|\mathbb{P}(|X_s - X_t| \leq u) - \int \mathbb{P}(X_s \in [a-t, a+t]) d\mathbb{P}(X_s = a)\right| \leq \phi(t-s).$$

Noting that

$$\begin{aligned}\int \mathbb{P}(X_s \in [a-u, a+u]) d\mathbb{P}(X_s = a) = &\int \mathbb{P}(X_s \in [a-u, a+u]) d\mathbb{P}(\tilde{X} = a) \\ = &\mathbb{P}(|X_1 - \tilde{X}| \leq u) = G(u),\end{aligned}$$

we have $\left|\mathbb{P}(|X_s - X_t| \leq u) - G(u)\right| \leq \phi(t-s)$. Hence, we have

$$\begin{aligned}\left|\mathbb{E}U_T(\phi_u) - G(u)\right| \leq &\frac{2}{T(T-1)} \sum_{1 \leq s < t \leq T} \left|\mathbb{P}(|X_s - X_t| \leq u) - G(u)\right| \\ \leq &\frac{2}{T(T-1)} \sum_{1 \leq s < t \leq T} \phi(t-s) \\ = &\frac{2}{T(T-1)} \sum_{k=1}^{T-1} (T-k)\phi(k) \leq \frac{2}{T} \sum_{k=1}^{\infty} \frac{1}{k^{1+\epsilon}}.\end{aligned}$$

Here the last inequality is due to $\phi(k) \leq 1/k^{1+\epsilon}$. This completes the proof. □

Lemma 4 provides the bias of $U_T(\psi_u)$ with respect to $G(u)$, which is the expectation

of $U_T(\psi_u)$ when the data points are independent. The bias increases with C_ϵ, which summarizes the degree of dependence over the process. Next, we proceed to the variance of $U_T(\psi_u)$.

Lemma 5. *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary ϕ-mixing process such that $\phi(n) \leq n^{-1-\epsilon}$ for any $n > 0$ and some constant $\epsilon > 0$, and $U_T(\psi_u)$ be defined in* (B.1)*. Then, for any $u > 0$, we have*

$$\mathbb{P}\{|U_T(\psi_u) - \mathbb{E}U_T(\psi_u)| \geq \tau\} \leq 2\exp\Big\{-\frac{T\tau^2}{2(1+2C_\epsilon)}\Big\}$$

for any $\tau > 0$, where $C_\epsilon = \sum_{k=1}^{\infty} 1/k^{1+\epsilon}$.

To prove Lemma 5, we first introduce a concentration inequality for ϕ-mixing processes.

Lemma 6. Kontorovich et al. (2008); Mohri and Rostamizadeh (2010) *Let $f : \Omega^T \to \mathbb{R}$ be a measurable function that is M-Lipschitz with respect to the Hamming metric for some $M > 0$:*

$$\sup_{x_1,\ldots,x_T,x_t'} \big|f(x_1,\ldots,x_t,\ldots,x_T) - f(x_1,\ldots,x_t',\ldots,x_T)\big| \leq M.$$

Then, for a stationary ϕ-mixing process $\{X_t\}_{t\in\mathbb{Z}}$, we have

$$\mathbb{P}\{|f(X_1,\ldots,X_T) - \mathbb{E}f(X_1,\ldots,X_T)| \geq \tau\} \leq 2\exp\Big[-\frac{2\tau^2}{M^2T\{1+2\sum_{k=1}^{T}\phi(k)\}}\Big].$$

for any $\tau > 0$.

Building on Lemma 6, we can proceed to the proof of Lemma 5.

Proof of Lemma 5. Let

$$f(x_1, \ldots, x_T) := TU_T(\psi_u) = \frac{2}{T-1} \sum_{s<t} I(|x_s - x_t| \leq u).$$

since replacing an element in $(x_1, \ldots, x_T)$, say, x_t, by x_t' only affects $T-1$ terms in the summation above, we have

$$\big|f(x_1, \ldots, x_t, \ldots, x_T) - f(x_1, \ldots, x_t', \ldots, x_T)\big| \leq 2.$$

Thus, by Lemma 6, we have

$$\mathbb{P}\{T|U_T(\psi_u) - \mathbb{E}U_T(\psi_u)| \geq \eta\} \leq 2\exp\Big[-\frac{\eta^2}{2T\{1 + 2\sum_{k=1}^{T}\phi(k)\}}\Big]$$

for any $\eta > 0$. Setting $\eta = T\tau$, we obtain

$$\begin{aligned}\mathbb{P}\{|U_T(\psi_u) - \mathbb{E}U_T(\psi_u)| \geq \tau\} \leq & 2\exp\Big[-\frac{T\tau^2}{2\{1 + 2\sum_{k=1}^{T}\phi(k)\}}\Big] \\ \leq & 2\exp\Big\{-\frac{T\tau^2}{2(1 + 2\sum_{k=1}^{\infty} 1/k^{1+\epsilon})}\Big\}.\end{aligned}$$

Here the last inequality is due to $\phi(k) \leq 1/k^{1+\epsilon}$. This completes the proof. □

Lemma 5 gives exponential tail probability for $U_T(\psi_u)$ around its expectation. Similar

to the bias of $U_T(\psi_u)$, the tail probability increases with C_ϵ. Thus, $U_T(\psi_u)$ is less concentrated around its expectation when the degree of dependence increases. Using Lemmas 4 and 5, we can derive the concentration inequality for $U_T(\psi_u)$ around $G(u)$.

Lemma 7. *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary ϕ-mixing process such that $\phi(n) \leq n^{-1-\epsilon}$ for any $n > 0$ and some constant $\epsilon > 0$. Suppose X_1 is absolutely continuous. Let $U_T(\psi_u)$ and $G(u)$ be defined as in Lemma 4. Then, for any $u > 0$, we have*

$$\mathbb{P}\{|U_T(\psi_u) - G(u)| \geq \tau\} \leq 2\exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\tau - \frac{2C_\epsilon}{T}\Big)^2\Big\}$$

for $\tau > 2C_\epsilon/T$ and $C_\epsilon = \sum_{k=1}^{\infty} 1/k^{1+\epsilon}$.

Proof. Using Lemma 4, we have

$$\begin{aligned}\mathbb{P}\{|U_T(\psi_u) - G(u)| \geq \tau\} \leq& \mathbb{P}\{|U_T(\psi_u) - \mathbb{E}U_T(\psi_u)| + |\mathbb{E}U_T(\psi_u) - G(u)| \geq \tau\} \\ \leq& \mathbb{P}\Big\{|U_T(\psi_u) - \mathbb{E}U_T(\psi_u)| \geq \tau - \frac{2C_\epsilon}{T}\Big\}.\end{aligned}$$

Applying Lemma 5 completes the proof. □

Now we can proceed to the concentration inequality of $\hat{\sigma}^{\mathrm{Q}}$.

Lemma 8. *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary ϕ-mixing process such that $\phi(n) \leq n^{-1-\epsilon}$ for any $n > 0$ and some constant $\epsilon > 0$. Let $\tilde{X}$ be an independent copy of X_1, and $q \in [0, 1]$ be an absolute constant. Suppose the following assumptions hold:*

1. *$Q(|X_1 - \tilde{X}|; q)$ and $\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q)$ are unique with probability 1.*

2. *There exist constants* $\kappa > 0$ *and* $\eta > 0$ *such that*

$$\inf_{|y-Q(|X_1-\tilde{X}|;q)|\leq\kappa} \frac{d}{dy}G(y) \geq \eta,$$

where G *is the distribution function of* $|X_1 - \tilde{X}|$.

Then, we have

$$\begin{aligned}&\mathbb{P}\big[|\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q) - Q(|X_1 - \tilde{X}|; q)| \geq u\big] \\ &\leq 2\exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\eta u - \frac{4C_\epsilon}{T}\Big)^2\Big\},\end{aligned} \tag{B.7}$$

when $4C_\epsilon/(\eta T) \leq u \leq \kappa$ *and* $C_\epsilon = \sum_{k=1}^{\infty} 1/k^{1+\epsilon}$.

Proof. We denote by G_T the empirical distribution function of $\{|X_s - X_t|\}_{1\leq s<t\leq T}$. G_T is non-decreasing and satisfies

$$q \leq G_T\big\{\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q)\big\} \leq q + \frac{2}{T(T-1)}.$$

The above inequality is because $\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q)$ is unique. Denote $G^{-1}(q) = Q(|X_1 - \tilde{X}|; q)$. Since $Q(|X_1 - \tilde{X}|; q)$ is unique, we have $G\{G^{-1}(q)\} = q$. Thus, we have

$$\begin{aligned}&\mathbb{P}\big[\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q) - Q(|X_1 - \tilde{X}|; q) \geq u\big] \\ &\leq \mathbb{P}\big[G_T\big\{\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q)\big\} \geq G_T\{G^{-1}(q) + u\}\big] \\ &\leq \mathbb{P}\big[q + \frac{2}{T(T-1)} \geq U_T\{\psi_{G^{-1}(q)+u}\}\big]\end{aligned}$$

$$= \mathbb{P}\big[-U_T\{\psi_{G^{-1}(q)+u}\} + G\{G^{-1}(q)+u\} \geq G\{G^{-1}(q)+u\} - q - \frac{2}{T(T-1)}\big],$$

where $U_T\{\psi_{G^{-1}(q)+u}\}$ is defined in Lemma 4. By Assumption 2, we have $G\{G^{-1}(q)+u\} - q \leq \eta$ when $u \leq \kappa$. Now, using Lemma 4, we have

$$\begin{aligned}
&\mathbb{P}\big[\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q) - Q(|X_1 - \tilde{X}|; q) \geq u\big] \\
&\leq \mathbb{P}\big[|U_T\{\psi_{G^{-1}(q)+u}\} - G\{G^{-1}(q)+u\}| \geq \eta u - \frac{2}{T(T-1)}\big] \\
&\leq 2\exp\Big[-\frac{T}{2(1+2C_\epsilon)}\Big\{\eta u - \frac{2}{T(T-1)} - \frac{2C_\epsilon}{T}\Big\}^2\Big] \\
&\leq 2\exp\Big[-\frac{T}{2(1+2C_\epsilon)}\Big\{\eta u - \frac{4C_\epsilon}{T}\Big\}^2\Big],
\end{aligned} \tag{B.8}$$

provided that $4C_\epsilon/(\eta T) \leq u \leq \kappa$. On the other hand, using the same technique, we have

$$\begin{aligned}
&\mathbb{P}\big[\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q) - Q(|X_1 - \tilde{X}|; q) \leq -u\big] \\
&\leq \mathbb{P}\big[G_T\big\{\hat{Q}(\{|X_s - X_t|\}_{1\leq s<t\leq T}; q)\big\} \leq G_T\{G^{-1}(q) - u\}\big] \\
&\leq \mathbb{P}\big[U_T\{\psi_{G^{-1}(q)+u}\} - G\{G^{-1}(q) - u\} \geq q - G\{G^{-1}(q) - u\}\big] \\
&\leq \mathbb{P}\big[|U_T\{\psi_{G^{-1}(q)+u}\} - G\{G^{-1}(q) - u\}| \geq \eta u\}\big] \\
&\leq 2\exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\eta u - \frac{2C_\epsilon}{T}\Big)^2\Big\},
\end{aligned} \tag{B.9}$$

provided that $2C_\epsilon/(\eta T) \leq u \leq \kappa$. Combining (B.8) and (B.9) completes the proof. □

Setting $q = 1/4$ in Lemma 8, we obtain the concentration inequality for $\hat{\sigma}^{\mathrm{Q}}$. Again, we

observe that the tail probability in (B.7) increases with C_ϵ, which represents the degree of serial dependence.

Now we have sufficient background for deriving the rate of convergence for $\hat{\mathbf{R}}^{\mathrm{Q}}$. Regarding $\tilde{\mathbf{R}}^{\mathrm{Q}}$, the following lemma connects its concentration probability with that of $\hat{\mathbf{R}}^{\mathrm{Q}}$.

Lemma 9. *For any $u > 0$, the solution $\tilde{\mathbf{R}}^{\mathrm{Q}}$ to the optimization problem* (3.4) *satisfies*

$$\mathbb{P}(\|\tilde{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq u) \leq \mathbb{P}\Big(\|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq \frac{u}{2}\Big),$$

provided that $\mathbf{R}^{\mathrm{Q}} \in S_\lambda$.

Proof. For any $u > 0$, we have

$$\mathbb{P}(\|\tilde{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq u) \leq \mathbb{P}(\|\tilde{\mathbf{R}}^{\mathrm{Q}} - \hat{\mathbf{R}}^{\mathrm{Q}}\|_{\max} + \|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq u)$$

Since $\mathbf{R}^{\mathrm{Q}}$ is feasible to (3.4), we have

$$\|\hat{\mathbf{R}}^{\mathrm{Q}} - \tilde{\mathbf{R}}^{\mathrm{Q}}\|_{\max} \leq \|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max}.$$

Combining the above two inequalities, we have

$$\mathbb{P}(\|\tilde{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq u) \leq \mathbb{P}(2\|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq u).$$

This completes the proof. □

B.2 Proofs of the Main Results

In this section, we provide technical proofs for the theoretical results.

B.2.1 Proof of Lemma 3

Proof. Since $\hat{\mathbf{w}}^{\mathrm{opt}}$ is feasible to (3.3), we have $R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) \geq R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}})$. Similarly, since $\mathbf{w}^{\mathrm{opt}}$ is feasible to (3.6), we have $R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}) \geq R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R})$. Thus, we have

$$
\begin{aligned}
&|R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) - R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}})| = R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) - R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) \\
=&R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) - R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}) + R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}) - R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}) + R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}) - R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) \\
\leq&R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) - R(\hat{\mathbf{w}}^{\mathrm{opt}}; \mathbf{R}) + R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}) - R(\mathbf{w}^{\mathrm{opt}}; \mathbf{R}^{\mathrm{Q}}) \\
\leq&2 \sup_{\|\mathbf{w}\|_1 \leq c} |R(\mathbf{w}; \mathbf{R}^{\mathrm{Q}}) - R(\mathbf{w}; \mathbf{R})| = 2 \sup_{\|\mathbf{w}\|_1 \leq c} |\mathbf{w}^{\mathsf{T}}(\mathbf{R}^{\mathrm{Q}} - \mathbf{R})\mathbf{w}| \leq 2c^2 \|\mathbf{R}^{\mathrm{Q}} - \mathbf{R}\|_{\max}.
\end{aligned}
$$

Here the last inequality is due to $|\mathbf{w}^{\mathsf{T}}(\mathbf{R}^{\mathrm{Q}} - \mathbf{R})\mathbf{w}| \leq \|\mathbf{w}\|_1^2 \|\mathbf{R}^{\mathrm{Q}} - \mathbf{R}\|_{\max}$. This completes the proof. □

B.2.2 Proof of Theorem 6

Proof. For notational brevity, we denote

$$
\begin{aligned}
&\hat{\sigma}_j^{\mathrm{Q}} := \hat{\sigma}^{\mathrm{Q}}(\{X_{tj}\}_{t=1}^T),\ \sigma_j^{\mathrm{Q}} := \sigma^{\mathrm{Q}}(X_j), \\
&\hat{\sigma}_{jk+}^{\mathrm{Q}} := \hat{\sigma}^{\mathrm{Q}}(\{X_{tj} + X_{tk}\}_{t=1}^T), \hat{\sigma}_{jk-}^{\mathrm{Q}} := \hat{\sigma}^{\mathrm{Q}}(\{X_{tj} - X_{tk}\}_{t=1}^T),
\end{aligned}
$$

$$\sigma^{\mathrm{Q}}_{jk+} := \sigma^{\mathrm{Q}}(X_{1j} + X_{1k}), \sigma^{\mathrm{Q}}_{jk-} := \sigma^{\mathrm{Q}}(X_{1j} - X_{1k}).$$

By definition, for any $u > 0$, we have

$$\begin{aligned}\mathbb{P}(|\hat{\mathbf{R}}^{\mathrm{Q}}_{jj} - \mathbf{R}^{\mathrm{Q}}_{jj}| \geq u) =& \mathbb{P}(|\hat{\sigma}^{\mathrm{Q}^2}_j - \sigma^{\mathrm{Q}^2}_j| \geq u) \leq \mathbb{P}\big((\hat{\sigma}^{\mathrm{Q}}_j - \sigma^{\mathrm{Q}}_j)^2 + 2\sigma^{\mathrm{Q}}_j|\hat{\sigma}^{\mathrm{Q}}_j - \sigma^{\mathrm{Q}}_j| \geq u\big) \\ \leq& \mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}}_j - \sigma^{\mathrm{Q}}_j| \geq \sqrt{\frac{u}{2}}\Big) + \mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}}_j - \sigma^{\mathrm{Q}}_j| \geq \frac{u}{4\sigma^{\mathrm{Q}}_j}\Big). \end{aligned} \tag{B.10}$$

The quantiles in the definitions of $\mathbf{R}^{\mathrm{Q}}$ and $\hat{\mathbf{R}}^{\mathrm{Q}}$ are unique due to Condition 2 and absolute continuity of $\boldsymbol{X}_1$. Hence, applying Lemma 8 and noting that $\sigma^{\mathrm{Q}}_j \leq \sigma^{\mathrm{Q}}_{\max}$, we have

$$\begin{aligned}&\mathbb{P}(|\hat{\mathbf{R}}^{\mathrm{Q}}_{jj} - \mathbf{R}^{\mathrm{Q}}_{jj}| \geq u) \leq \\ &2\exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\eta\sqrt{\frac{u}{2}} - \frac{4C_\epsilon}{T}\Big)^2\Big\} + 2\exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\frac{\eta u}{4\sigma^{\mathrm{Q}}_{\max}} - \frac{4C_\epsilon}{T}\Big)^2\Big\}, \end{aligned} \tag{B.11}$$

when $4C_\epsilon/(\eta T) \leq \sqrt{u/2}, u/(4\sigma^{\mathrm{Q}}_{\max}) \leq \kappa$. Now, for the off-diagonal entries, we have

$$\begin{aligned}\mathbb{P}(|\hat{\mathbf{R}}^{\mathrm{Q}}_{jk} - \mathbf{R}^{\mathrm{Q}}_{jk}| \geq u) \leq& \mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}^2}_{jk+} - \sigma^{\mathrm{Q}^2}_{jk+}| + |\hat{\sigma}^{\mathrm{Q}^2}_{jk-} - \sigma^{\mathrm{Q}^2}_{jk-}| \geq 4u\Big) \\ \leq& \mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}^2}_{jk+} - \sigma^{\mathrm{Q}^2}_{jk+}| \geq 2u\Big) + \mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}^2}_{jk-} - \sigma^{\mathrm{Q}^2}_{jk-}| \geq 2u\Big). \end{aligned}$$

Using the same technique as in (B.10), we further have

$$\mathbb{P}(|\hat{\mathbf{R}}^{\mathrm{Q}}_{jk} - \mathbf{R}^{\mathrm{Q}}_{jk}| \geq u) \leq \mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}}_{jk+} - \sigma^{\mathrm{Q}}_{jk+}| \geq \sqrt{u}\Big) + \mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}}_{jk+} - \sigma^{\mathrm{Q}}_{jk+}| \geq \frac{u}{2\sigma^{\mathrm{Q}}_{jk+}}\Big) +$$

$$\mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}}_{jk-} - \sigma^{\mathrm{Q}}_{jk-}| \geq \sqrt{u}\Big) + \mathbb{P}\Big(|\hat{\sigma}^{\mathrm{Q}}_{jk-} - \sigma^{\mathrm{Q}}_{jk-}| \geq \frac{u}{2\sigma^{\mathrm{Q}}_{jk-}}\Big).$$

Applying Lemma 8 and noting that $\sigma^{\mathrm{Q}}_{jk+} \leq \sigma^{\mathrm{Q}}_{\max}$ and $\sigma^{\mathrm{Q}}_{jk-} \leq \sigma^{\mathrm{Q}}_{\max}$, we have

$$\begin{aligned}
&\mathbb{P}(|\hat{\mathbf{R}}^{\mathrm{Q}}_{jk} - \mathbf{R}^{\mathrm{Q}}_{jk}| \geq u) \leq \\
&4\exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\eta\sqrt{u} - \frac{4C_\epsilon}{T}\Big)^2\Big\} + 4\exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\frac{\eta u}{2\sigma^{\mathrm{Q}}_{\max}} - \frac{4C_\epsilon}{T}\Big)^2\Big\}, \quad \text{(B.12)}
\end{aligned}$$

when $4C_\epsilon/(\eta T) \leq \sqrt{u}, u/(2\sigma^{\mathrm{Q}}_{\max}) \leq \kappa$. Combining (B.11) and (B.12), we obtain

$$\begin{aligned}
&\mathbb{P}(\|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq u) \leq \sum_{j,k=1}^{d} \mathbb{P}(|\hat{\mathbf{R}}^{\mathrm{Q}}_{jk} - \mathbf{R}^{\mathrm{Q}}_{jk}| \geq u) \\
&\leq 4d^2\Big[\exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\eta\sqrt{\frac{u}{2}} - \frac{4C_\epsilon}{T}\Big)^2\Big\} + \exp\Big\{-\frac{T}{2(1+2C_\epsilon)}\Big(\frac{\eta u}{4\sigma^{\mathrm{Q}}_{\max}} - \frac{4C_\epsilon}{T}\Big)^2\Big\}\Big] \\
&\leq 8\max\Big\{\underbrace{d^2\exp\Big[-\frac{T}{2(1+2C_\epsilon)}\Big(\eta\sqrt{\frac{u}{2}} - \frac{4C_\epsilon}{T}\Big)^2\Big]}_{A_1(u)}, \underbrace{d^2\exp\Big[-\frac{T}{2(1+2C_\epsilon)}\Big(\frac{\eta u}{4\sigma^{\mathrm{Q}}_{\max}} - \frac{4C_\epsilon}{T}\Big)^2\Big]}_{A_2(u)}\Big\},
\end{aligned}$$

when we have

$$4C_\epsilon/(\eta T) \leq \sqrt{\frac{u}{2}}, \sqrt{u}, \frac{u}{4\sigma^{\mathrm{Q}}_{\max}}, \frac{u}{2\sigma^{\mathrm{Q}}_{\max}} \leq \kappa. \quad \text{(B.13)}$$

Setting $A_1(u_1) = \alpha^2$, we obtain

$$u_1 = \frac{2}{\eta^2}\Big[\sqrt{\frac{4(1+2C_\epsilon)(\log d - \log\alpha)}{T}} + \frac{4C_\epsilon}{T}\Big]^2.$$

Setting $A_2(u_2) = \alpha^2$, we obtain

$$u_2 = \frac{4\sigma_{\max}^{\mathrm{Q}}}{\eta}\Big[\sqrt{\frac{4(1+2C_\epsilon)(\log d - \log \alpha)}{T}} + \frac{4C_\epsilon}{T}\Big].$$

Now set $u = r_T = \max(u_1, u_2)$. (B.13) is satisfied when T is large enough. If $u_1 \geq u_2$, since $A_2(u)$ is a non-increasing function of u, we have $A_2(u_1) \leq A_2(u_2) = \alpha^2$. Thus, we have

$$\mathbb{P}(\|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq r_T) \leq 8 \max\{A_1(u), A_2(u)\} \leq 8\alpha^2.$$

On the other hand, if $u_1 < u_2$, we have $r_T = u_2$. Since $A_1(u)$ is a non-increasing function of u, we have $A_1(u_2) \leq A_1(u_1) = \alpha^2$. Thus, we still have

$$\mathbb{P}(\|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq r_T) \leq 8 \max\{A_1(u), A_2(u)\} \leq 8\alpha^2.$$

This proves (3.7). Applying Lemma 9, we have

$$\mathbb{P}(\|\tilde{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq 2r_T) \leq \mathbb{P}(\|\hat{\mathbf{R}}^{\mathrm{Q}} - \mathbf{R}^{\mathrm{Q}}\|_{\max} \geq r_T) \leq 8\alpha^2.$$

This proves (3.9). □

B.2.3 Proof of Theorem 9

Proof. We will utilize an equivalent definition of elliptical distributions. Specifically, $\boldsymbol{X}$ is elliptically distributed with location $\boldsymbol{\mu}$ and scatter $\mathbf{S}$ if and only if the characteristic function of $\boldsymbol{X}$ is $\psi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\mathsf{T}\boldsymbol{\mu})\varphi(\boldsymbol{t}^\mathsf{T}\mathbf{S}\boldsymbol{t})$ for some function φ Fang et al. (1990). Since $\tilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$, the characteristic function of $\boldsymbol{X} - \tilde{\boldsymbol{X}}$ is

$$\psi_{\boldsymbol{X}-\tilde{\boldsymbol{X}}}(\boldsymbol{t}) = \mathbb{E}\exp\{i\boldsymbol{t}^\mathsf{T}(\boldsymbol{X} - \tilde{\boldsymbol{X}})\} = \mathbb{E}\exp(i\boldsymbol{t}^\mathsf{T}\boldsymbol{X})\mathbb{E}\exp(-i\boldsymbol{t}^\mathsf{T}\tilde{\boldsymbol{X}}) = \varphi(\boldsymbol{t}^\mathsf{T}\mathbf{S}\boldsymbol{t})^2.$$

Thus, $\boldsymbol{X} - \tilde{\boldsymbol{X}} \sim \mathrm{EC}_d(\mathbf{0}, \mathbf{S}, \zeta)$ is elliptical distributed for some generating variate ζ. By Theorem 2.6 in Fang et al. (1990), we have

$$X_j - \tilde{X}_j \sim \mathrm{EC}_1(0, \mathbf{S}_{jj}, \sqrt{D}\zeta),$$

where $D \sim \mathrm{Beta}(1/2, (d-1)/2)$ follows a Beta distribution. Since $\boldsymbol{X}$ is absolutely continuous, we have $\mathbf{S}_{jj} > 0$. Thus, we have

$$\begin{aligned}\mathbf{R}^{\mathrm{Q}}_{jj} &= Q(|X_j - \tilde{X}_j|; 1/4)^2 = Q\{(X_j - \tilde{X}_j)^2; 1/4\} \\ &= \mathbf{S}_{jj} Q\Big\{\frac{(X_j - \tilde{X}_j)^2}{\mathbf{S}_{jj}}; \frac{1}{4}\Big\} = \mathbf{S}_{jj} Q(D\zeta^2; 1/4). \end{aligned} \tag{B.14}$$

By Theorems 2.15 and 2.16 in Fang et al. (1990), we have, for $j \neq k$,

$$X_j + X_k - \tilde{X}_j - \tilde{X}_k \sim \mathrm{EC}_1(0, \mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk}, \sqrt{D}\zeta),$$

$$X_j - X_k - \tilde{X}_j + \tilde{X}_k \sim \mathrm{EC}_1(0, \mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk}, \sqrt{D}\zeta).$$

Thus, if $\mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk} > 0$ and $\mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk} > 0$, we have

$$\begin{aligned}
\sigma^{\mathrm{Q}}(X_j + X_k)^2 =& Q(|X_j + X_k - \tilde{X}_j - \tilde{X}_k|; 1/4)^2 \\
=& (\mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk}) Q\Big\{ \frac{(X_j + X_k - \tilde{X}_j - \tilde{X}_k)^2}{\mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk}}; \frac{1}{4} \Big\} \\
=& (\mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk}) Q(D\zeta^2; 1/4); \qquad \text{(B.15)} \\
\sigma^{\mathrm{Q}}(X_j - X_k)^2 =& Q(|X_j - X_k - \tilde{X}_j + \tilde{X}_k|; 1/4)^2 \\
=& (\mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk}) Q\Big\{ \frac{(X_j - X_k - \tilde{X}_j + \tilde{X}_k)^2}{\mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk}}; \frac{1}{4} \Big\} \\
=& (\mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk}) Q(D\zeta^2; 1/4). \qquad \text{(B.16)}
\end{aligned}$$

Note that when $\mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk} = 0$ or $\mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk} = 0$, we have $\sigma^{\mathrm{Q}}(X_j + X_k) = 0$ or $\sigma^{\mathrm{Q}}(X_j - X_k) = 0$. Thus, we still have $\sigma^{\mathrm{Q}}(X_j + X_k)^2 = (\mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk}) Q(D\zeta^2; 1/4)$ and $\sigma^{\mathrm{Q}}(X_j - X_k)^2 = (\mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk}) Q(D\zeta^2; 1/4)$. Thus, we have

$$\mathbf{R}^{\mathrm{Q}}_{jk} = \frac{1}{4}\{\sigma^{\mathrm{Q}}(X_j + X_k)^2 - \sigma^{\mathrm{Q}}(X_j - X_k)^2\} = \mathbf{S}_{jk} Q(D\zeta^2; 1/4). \qquad \text{(B.17)}$$

Combining (B.14) and (B.17), we have (3.11) with $m^{\mathrm{Q}} = Q(D\zeta^2; 1/4)$.

When $0 < \mathbb{E}\xi^2 < \infty$, by the the corollary on Page 34 in Fang et al. (1990), we have

$\mathbf{S} = r\boldsymbol{\Sigma}/\mathbb{E}\xi^2$, where $r = \mathrm{rank}(\mathbf{S})$. Thus, we have

$$\mathbf{R}^{\mathrm{Q}} = Q(D\zeta^2; 1/4)\mathbf{S} = \frac{r}{\mathbb{E}\xi^2} Q(D\zeta^2; 1/4)\boldsymbol{\Sigma} = c^{\mathrm{Q}}\boldsymbol{\Sigma}.$$

This proves (3.12). By (B.14), we have

$$c^{\mathrm{Q}} = Q\Big\{\frac{r(X_j - \tilde{X}_j)^2}{\mathbb{E}\xi^2 \mathbf{S}_{jj}}; \frac{1}{4}\Big\} = Q\Big\{\frac{(X_j - \tilde{X}_j)^2}{\mathrm{Var}(X_j)}; \frac{1}{4}\Big\}.$$

The last equality is due to $\mathbf{S} = r\boldsymbol{\Sigma}/\mathbb{E}\xi^2$. Similarly, when $\mathrm{Var}(X_j + X_k) > 0$ and $\mathrm{Var}(X_j - X_k) > 0$, by (B.15) and (B.16), we have

$$\begin{aligned} c^{\mathrm{Q}} =& Q\Big\{\frac{r(X_j + X_k - \tilde{X}_j - \tilde{X}_k)^2}{\mathbb{E}\xi^2(\mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk}}); \frac{1}{4}\Big\} = Q\Big\{\frac{(X_j + X_k - \tilde{X}_j - \tilde{X}_k)^2}{\mathrm{Var}(X_j + X_k)}; \frac{1}{4}\Big\}; \\ c^{\mathrm{Q}} =& Q\Big\{\frac{r(X_j - X_k - \tilde{X}_j + \tilde{X}_k)^2}{\mathbb{E}\xi^2(\mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk}}); \frac{1}{4}\Big\} = Q\Big\{\frac{(X_j - X_k - \tilde{X}_j + \tilde{X}_k)^2}{\mathrm{Var}(X_j - X_k)}; \frac{1}{4}\Big\}. \end{aligned}$$

This proves (3.13). □

B.3 Matrix Projection

In this section, we summarize the algorithm proposed in Xu and Shao (2012b) for solving the matrix projection problem (3.4). Let

$$\Omega_1 := \Big\{\boldsymbol{x} = \mathrm{vec}(\mathbf{X}) : \mathbf{X} \in S_\lambda\Big\}$$

$$\Omega_2 := \Big\{ \boldsymbol{z} = \mathrm{vec}(\mathbf{Z}) : \mathbf{Z} \in \mathbb{R}^{d\times d}, \mathbf{Z} = \mathbf{Z}^\mathsf{T}, \sum_{i,j=1}^{d} |Z_{ij}| \leq 1 \Big\}.$$

For any symmetric matrix $\mathbf{V} \in \mathbb{R}^{d\times d}$ and $\mathbf{v} = \mathrm{vec}(\mathbf{V})$, define the projection of $\mathbf{v}$ onto Ω_i as

$$P_{\Omega_i}(\mathbf{v}) = \arg\min_{\boldsymbol{x}\in\Omega_i} \Big\| \boldsymbol{x} - \mathbf{v} \Big\|_2^2, \tag{B.18}$$

for $i = 1, 2$. The algorithm for solving (3.4) builds on solutions to the problems in (B.18). Solving for $P_{\Omega_1}(\mathbf{v})$ is straightforward. It's well known that

$$P_{\Omega_1}(\mathbf{v}) = \mathrm{vec}(\mathbf{U}\tilde{\mathbf{\Lambda}}\mathbf{U}^\mathsf{T}), \tag{B.19}$$

where $\mathbf{V} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ is a spectral decomposition of $\mathbf{V}$, $\tilde{\mathbf{\Lambda}} = \mathrm{diag}(\tilde{\mathbf{\Lambda}}_{11}, \ldots, \tilde{\mathbf{\Lambda}}_{dd})$ and $\tilde{\mathbf{\Lambda}}_{ii} = \min\Big\{ \max\{\mathbf{\Lambda}_{ii}, \lambda_{\min}\}, \lambda_{\max} \Big\}$ for $i = 1, \ldots, d$.

Next we solve for $P_{\Omega_2}(\mathbf{v})$. Let $\mathrm{sign}(\mathbf{v}) = \{\mathrm{sign}(v_1), \ldots, \mathrm{sign}(v_d)\}^\mathsf{T}$ be a vector of the signs of $\mathbf{v}$'s entries. Denote $|\mathbf{v}| = \mathrm{sign}(\mathbf{v}) \circ \mathbf{v}$ and $\tilde{\mathbf{v}} = T_{|\mathbf{v}|}(|\mathbf{v}|)$, where $T_{|\mathbf{v}|}$ is a permutation transformation that sorts the elements of $|\mathbf{v}|$ in descending order. Now, if $\mathbf{1}^\mathsf{T}\tilde{\mathbf{v}} \leq 1$, we set $(\tilde{\boldsymbol{x}}, \tilde{y}) = (\tilde{\mathbf{v}}, 0)$. If $\mathbf{1}^\mathsf{T}\tilde{\mathbf{v}} > 1$, let $\Delta\mathbf{v} := (\tilde{v}_1 - \tilde{v}_2, \ldots, \tilde{v}_{d-1} - \tilde{v}_d, \tilde{v}_d)^\mathsf{T} \in \mathbb{R}^d$. Note that $\Delta v_i \geq 0$ for $i = 1, \ldots, d$ and $\sum_{i=1}^{d} i\Delta v_i = \mathbf{1}^\mathsf{T}\tilde{\mathbf{v}} > 1$. Thus, there exists a smallest integer

Algorithm 1 Solving matrix projection problem (3.4)

$\tilde{\mathbf{R}}^{\mathrm{Q}} \leftarrow$ MatrixProjection($\hat{\mathbf{R}}^{\mathrm{Q}}$, $\lambda_{\min}$, $\lambda_{\max}$, $\boldsymbol{x}^0$, $\boldsymbol{z}^0$, γ, ϵ, N)
$\boldsymbol{r} \leftarrow \mathrm{vec}(\hat{\mathbf{R}}^{\mathrm{Q}})$
for $k = 0, \ldots, N$ **do**
  $\boldsymbol{e}_x^k \leftarrow \boldsymbol{x}^k - P_{\Omega_1}(\boldsymbol{x}^k - \boldsymbol{z}^k)$
  $\boldsymbol{e}_z^k \leftarrow \boldsymbol{z}^k - P_{\Omega_2}(\boldsymbol{z}^k + \boldsymbol{x}^k - \boldsymbol{r})$
  $\boldsymbol{e}^k \leftarrow (\boldsymbol{e}_x^k, \boldsymbol{e}_z^k)^{\mathsf{T}}$
  if $\|\boldsymbol{e}^k\|_{\max} < \epsilon$, **then**
    break
  else
    $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k - \gamma(\boldsymbol{e}_x^k - \boldsymbol{e}_z^k)/2$
    $\boldsymbol{z}^{k+1} \leftarrow \boldsymbol{z}^k - \gamma(\boldsymbol{e}_x^k + \boldsymbol{e}_z^k)/2$
  end if
end for
return $\tilde{\mathbf{R}}^{\mathrm{Q}} = \mathrm{mat}(\boldsymbol{x}^k)$

K such that $\sum_{i=1}^{K} i\Delta v_i \geq 1$. In this case, we set

$$\tilde{y} = \frac{1}{K}\Big(\sum_{i=1}^{K} \tilde{v}_i - 1\Big) \text{ and } \tilde{\boldsymbol{x}} = (\tilde{v}_1 - \tilde{y}, \ldots, \tilde{v}_K - \tilde{y}, 0, \ldots, 0)^{\mathsf{T}} \in \mathbb{R}^d.$$

Now we can express $P_{\Omega_2}(\mathbf{v})$ as

$$P_{\Omega_2}(\mathbf{v}) = \mathrm{sign}(\mathbf{v}) \circ T_{|\mathbf{v}|}^{-1}(\tilde{\boldsymbol{x}}). \tag{B.20}$$

Next we solve the matrix projection problem in (3.4). Recall that $\hat{\mathbf{R}}^{\mathrm{Q}}$ is the matrix to be projected to S_λ. Since for any vector $\boldsymbol{y} \in \mathbb{R}^d$, we have $\|\boldsymbol{y}\|_{\max} = \max_{\boldsymbol{c}\in\mathbb{R}^d, \|c\|_1 \leq 1} \boldsymbol{c}^{\mathsf{T}}\boldsymbol{y}$, it follows that problem (3.4) can be reformulated as the following mini-max problem:

$$\min_{\boldsymbol{x}\in\Omega_1} \max_{\boldsymbol{z}\in\Omega_2} \boldsymbol{z}^{\mathsf{T}}\Big\{\boldsymbol{x} - \mathrm{vec}\big(\hat{\mathbf{R}}^{\mathrm{Q}}\big)\Big\}. \tag{B.21}$$

If $(\boldsymbol{x}^{\mathrm{opt}},\ \boldsymbol{z}^{\mathrm{opt}})$ is a solution to problem (B.21), then $\mathrm{mat}(\boldsymbol{x}^{\mathrm{opt}})$ is a solution to problem (3.4). Algorithm 1 gives the pseudo code for solving problem (B.21), and thus (3.4). Recall that $0 \leq \lambda_{\min} < \lambda_{\max} \leq \infty$ are the lower and upper bounds of the eigenvalues of the projection. $\boldsymbol{x}^0 \in \Omega_1$ and $\boldsymbol{z}^0 \in \Omega_2$ are arbitrary initial points. $\gamma \in (0, 2)$ is a parameter controlling the step lengths of every iteration. $\epsilon > 0$ is a prespecified tolerance level. $N \in \mathbb{N}$ is the maximum number of iterations desired. The convergence of Algorithm 1 is guaranteed by the following theorem.

Theorem 24 (Xu and Shao (2012b))**.** *Let* $\boldsymbol{u}^{\mathrm{opt}} := (\boldsymbol{x}^{\mathrm{opt}}, \boldsymbol{z}^{\mathrm{opt}})$ *be a solution to* (B.21)*. Denote* $\boldsymbol{u}^k := (\boldsymbol{x}^{k\mathsf{T}}, \boldsymbol{z}^{k\mathsf{T}})^{\mathsf{T}}$ *and* $\boldsymbol{e}_u^k := (\boldsymbol{e}_x^{k\mathsf{T}}, \boldsymbol{e}_z^{k\mathsf{T}})^{\mathsf{T}}$*. Then Algorithm 1 produces a sequence* $\{\boldsymbol{u}^k\}$ *satisfying*

$$\left\|\boldsymbol{u}^{k+1} - \boldsymbol{u}^{\mathrm{opt}}\right\|^2 \leq \left\|\boldsymbol{u}^k - \boldsymbol{u}^{\mathrm{opt}}\right\|^2 + \frac{\gamma(2-\gamma)}{2}\left\|\boldsymbol{e}_u^k\right\|^2.$$

Appendix C

Appendix to Chapter 4

C.1 Concentration Inequalities under Weak Dependence

In this section, we develop a concentration inequality for sums of weakly dependent random variables. We first reformulate Theorem 1 in Doukhan and Neumann (2007).

Lemma 10. *Suppose $X_1, \ldots, X_T$ are real-valued random variables with mean 0, defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Let $\Psi : \mathbb{N}^2 \to \mathbb{N}$ be one of the four functions defined in Condition 3. Assume that there exist constants $K, M, L_1, L_2 > 0$, $a, b \geq 0$, and a nonincreasing sequence of real coefficients $\{\rho(t)\}_{t\geq 0}$ such that for any u-tuple $(s_1, \ldots, s_u)$*

and v-tuple $(t_1, \ldots, t_v)$ with $1 \leq s_1 \leq \cdots \leq s_u < t_1 \leq \cdots \leq t_v \leq T$, we have

$$\Big|\mathrm{Cov}\Big(\prod_{t=1}^{u} X_{s_t}, \prod_{j=1}^{v} X_{t_j}\Big)\Big| \leq K^2 M^{u+v}\{(u+v)!\}^b \Psi(u,v)\rho(t_1 - s_u), \tag{C.1}$$

where the sequence $\{\rho(t)\}_{t\geq 0}$ satisfies

$$\sum_{n=0}^{\infty}(n+1)^k \rho(n) \leq L_1 L_2^k (k!)^a, \text{ for any } k \geq 0 \text{ and } k \in \mathbb{Z}. \tag{C.2}$$

Moreover, we require that the following moment condition holds:

$$\mathbb{E}|X_t|^k \leq (k!)^b M^k, \ t = 1, \ldots, T, \text{ for any } k \geq 0 \text{ and } k \in \mathbb{Z}. \tag{C.3}$$

Then, for $S_T := \sum_{t=1}^{T} X_t$ and any $t > 0$, we have

$$\mathbb{P}(S_T \geq t) \leq \exp\Big\{-\frac{t^2}{C_1 T + C_2 t^{(2a+2b+3)/(a+b+2)}}\Big\},$$

where C_1 and C_2 are constants given by

$$C_1 = 2^{a+b+3} K^2 M^2 L_1 (K^2 \vee 2) \text{ and } C_2 = 2\{M L_2 (K^2 \vee 2)\}^{1/(a+b+2)}.$$

Proof. The proof follows that of Theorem 1 in Doukhan and Neumann (2007) with minor modifications, as listed below. We inherit the notation in Doukhan and Neumann (2007).

Equation (30) in Doukhan and Neumann (2007) can be strengthened to

$$\mathbb{E}|Y_j| \leq 2^{k-j-1}\{(k-j+1)!\}^b K^2 M^k \rho(t_{t+1}-t_t).$$

This leads to

$$|\overline{\mathbb{E}}(X_{t_1}\cdots X_{t_k})| \leq 2^{k-1}(k!)^b k^2 M^k \rho(t_{t+1}-t_t), \tag{C.4}$$

which corresponds to Lemma 13 in Doukhan and Neumann (2007). Using (C.4), we obtain that

$$\begin{aligned}\left|\Gamma(X_{t_1},\ldots,X_{t_k})\right| \leq & \sum_{\nu=1}^{k}\sum_{\bigcup_{p=1}^{\nu} I_p = I} N_\nu(I_1,\ldots,I_\nu) 2^{k-\nu}(k!)^b K^{2\nu} M^k \min_{1\leq t<k}\rho(t_{t+1}-t_t) \\ \leq & K^2(K^2\vee 2)^{k-1} M^k (k!)^b\{(k-1)!\}\min_{1\leq t<k}\rho(t_{t+1}-t_t).\end{aligned}$$

Thus, we have

$$\left|\Gamma_k(S_T)\right| \leq nK^2(K^2\vee 2)^{k-1}M^k(k!)^{b+1}\sum_{s=0}^{T-1}(s+1)^{k-2}\rho(s). \tag{C.5}$$

Equation (C.5) corresponds to Lemma 14 in Doukhan and Neumann (2007). The rest follows the same technique as in Doukhan and Neumann (2007). □

Equations (C.1) and (C.2) characterize the dependence structure of the sequence $X_1, \ldots, X_T$. In detail, the covariance between two blocks of observations converges to 0 as the gap

between the blocks increases. (C.2) specifies the speed of the convergence. Equation (C.3) is a moment condition. In the next lemma, we further show that these conditions are location and scale invariant.

Lemma 11. *Let $X_1, \ldots, X_T$ be a sequence of random variables satisfying* (C.1)-(C.3). *Let $\{\mu_t\}_{t=1}^T$ and $\{\gamma_t\}_{t=1}^T$ be uniformly bounded real sequences in the sense that $|\mu_t| \leq \mu$, $0 < \gamma_t \leq \gamma$, $t = 1, \ldots, T$, where μ and γ are constants. Let $Y_1, \ldots, Y_T$ be a location-scale transformed sequence defined as*

$$Y_t := \gamma_t(X_t + \mu_t),\ t = 1 \ldots, T.$$

Then (C.1)-(C.3) *are satisfied by $Y_1, \ldots, Y_T$ with M replaced by $\gamma(M + \mu)$.*

Proof. Equation (C.3) can be easily verified for $Y_1, \ldots, Y_T$:

$$\mathbb{E}|Y_t|^k = \mathbb{E}\Big|\gamma_t(X_t + \mu_t)\Big|^k \leq \gamma^k \sum_{j=0}^{k} \mathbb{E}|X_t|^j |\mu_t|^{k-j} \leq (k!)^b \Big\{\gamma(M + \mu)\Big\}^k.$$

The last inequality follows from (C.3). Next, we verify that $Y_1, \ldots, Y_T$ also satisfy (C.1) and (C.2). Let $\mathcal{S} := \{s_1, \ldots, s_u\}$, $\mathcal{T} := \{t_1, \ldots, t_v\}$, and $\mathcal{R} := \mathcal{S} \cup \mathcal{T}$. By the definiton of $Y_1, \ldots, Y_T$, we have

$$\begin{aligned}\mathbb{E}\prod_{t\in\mathcal{R}} Y_t &= \prod_{t\in\mathcal{R}} \gamma_t \mathbb{E}\prod_{j\in\mathcal{R}}(X_j + \mu_j) = \prod_{t\in\mathcal{R}} \gamma_t \sum_{\mathcal{U}\subseteq\mathcal{R}} \prod_{j\in\mathcal{R}\setminus\mathcal{U}} \mu_j \mathbb{E}\prod_{k\in\mathcal{U}} X_k \\ &= \prod_{t\in\mathcal{R}} \gamma_t \sum_{\mathcal{U}\subseteq\mathcal{S},\mathcal{V}\subseteq\mathcal{T}} \prod_{j\in\mathcal{R}\setminus(\mathcal{U}\cup\mathcal{V})} \mu_j \mathbb{E}\prod_{k\in\mathcal{U}\cup\mathcal{V}} X_k. \qquad \text{(C.6)}\end{aligned}$$

Applying the same derivation on $\mathbb{E}\prod_{t\in\mathcal{S}} Y_t$ and $\mathbb{E}\prod_{j\in\mathcal{T}} Y_j$, we obtain

$$\begin{aligned}\mathbb{E}\prod_{t\in\mathcal{S}} Y_t \mathbb{E}\prod_{j\in\mathcal{T}} Y_j =& \prod_{t\in\mathcal{R}} \gamma_t \Big(\sum_{\mathcal{U}\subseteq\mathcal{S}} \prod_{j\in\mathcal{S}\setminus\mathcal{U}} \mu_j \mathbb{E}\prod_{k\in\mathcal{U}} X_k\Big)\Big(\sum_{\mathcal{V}\subseteq\mathcal{T}} \prod_{j\in\mathcal{T}\setminus\mathcal{V}} \mu_j \mathbb{E}\prod_{k\in\mathcal{V}} X_k\Big) \\ =& \prod_{t\in\mathcal{R}} \gamma_t \sum_{\mathcal{U}\subseteq\mathcal{S},\mathcal{V}\subseteq\mathcal{T}} \prod_{j\in\mathcal{R}\setminus(\mathcal{U}\cup\mathcal{V})} \mu_j \mathbb{E}\prod_{k\in\mathcal{U}} X_k \mathbb{E}\prod_{\ell\in\mathcal{V}} X_l. \end{aligned} \tag{C.7}$$

By the definition of covariance, we have

$$\Big|\mathrm{Cov}(\prod_{t\in\mathcal{S}} Y_t, \prod_{t\in\mathcal{T}} Y_t)\Big| = \Big|\mathbb{E}\prod_{t\in\mathcal{R}} Y_t - \mathbb{E}\prod_{j\in\mathcal{S}} Y_j \mathbb{E}\prod_{k\in\mathcal{T}} Y_k\Big|.$$

Plugging (C.6) and (C.7) into the above equation, we have

$$\begin{aligned}\Big|\mathrm{Cov}(\prod_{t\in\mathcal{S}} Y_t, \prod_{t\in\mathcal{T}} Y_t)\Big| =& \Big|\prod_{t\in\mathcal{R}} \gamma_t \Big\{\sum_{\mathcal{U}\subseteq\mathcal{S},\mathcal{V}\subseteq\mathcal{T}} \prod_{j\in\mathcal{R}\setminus(\mathcal{U}\cup\mathcal{V})} \mu_j \Big(\mathbb{E}\prod_{k\in\mathcal{U}\cup\mathcal{V}} X_k - \mathbb{E}\prod_{\ell\in\mathcal{U}} X_l \mathbb{E}\prod_{m\in\mathcal{V}} X_m\Big)\Big\}\Big| \\ \le& \prod_{t\in\mathcal{R}} \gamma_t \Big\{\sum_{\mathcal{U}\subseteq\mathcal{S},\mathcal{V}\subseteq\mathcal{T}} \prod_{j\in\mathcal{R}\setminus(\mathcal{U}\cup\mathcal{V})} \mu_j \Big|\mathbb{E}\prod_{k\in\mathcal{U}\cup\mathcal{V}} X_k - \mathbb{E}\prod_{\ell\in\mathcal{U}} X_l \mathbb{E}\prod_{m\in\mathcal{V}} X_m\Big|\Big\} \\ =& \prod_{t\in\mathcal{R}} \gamma_t \Big\{\sum_{\mathcal{U}\subseteq\mathcal{S},\mathcal{V}\subseteq\mathcal{T}} \prod_{j\in\mathcal{R}\setminus(\mathcal{U}\cup\mathcal{V})} \mu_j \Big|\mathrm{Cov}(\prod_{k\in\mathcal{U}} X_k, \prod_{\ell\in\mathcal{V}} X_l)\Big|\Big\}. \end{aligned} \tag{C.8}$$

Now, (C.1) for $X_1, \ldots, X_T$ implies that

$$\begin{aligned}\Big|\mathrm{Cov}(\prod_{t\in\mathcal{U}} X_t, \prod_{t\in\mathcal{V}} X_t)\Big| \le& K^2 M^{|\mathcal{U}|+|\mathcal{V}|}\Big\{(|\mathcal{U}|+|\mathcal{V}|)!\Big\}^b \Psi(|\mathcal{U}|,|\mathcal{V}|)\rho\Big\{d(\mathcal{U},\mathcal{V})\Big\} \\ \le& K^2 M^{|\mathcal{U}|+|\mathcal{V}|}\Big\{(u+v)!\Big\}^b \Psi(u,v)\rho(t_1 - s_u), \end{aligned} \tag{C.9}$$

where the last inequality is due to $\mathcal{U} \subseteq \mathcal{S}$ and $\mathcal{V} \subseteq \mathcal{T}$. Plugging (C.9) into (C.8), we have

$$\begin{aligned}\Big|\mathrm{Cov}(\prod_{t\in\mathcal{S}} Y_t, \prod_{t\in\mathcal{T}} Y_t)\Big| \leq & \prod_{t\in\mathcal{R}} \gamma_t \Big\{ K^2 \Big\{ (u+v)! \Big\}^b \Psi(u,v)\rho(t_1 - s_u) \sum_{\mathcal{U}\subseteq\mathcal{S},\mathcal{V}\subseteq\mathcal{T}} M^{|\mathcal{U}|+|\mathcal{V}|} \prod_{j\in\mathcal{R}\setminus(\mathcal{U}\cup\mathcal{V})} \mu_j \Big\} \\ = & \prod_{t\in\mathcal{R}} \gamma_t K^2 \Big\{ (u+v)! \Big\}^b \Psi(u,v)\rho(t_1 - s_u) \Big(\sum_{\mathcal{W}\subseteq\mathcal{R}} M^{|\mathcal{W}|} \prod_{j\in\mathcal{R}\setminus\mathcal{W}} \mu_j \Big).\end{aligned}$$

Noting that $\sum_{\mathcal{W}\subseteq\mathcal{R}} M^{|\mathcal{W}|} \prod_{j\in\mathcal{R}\setminus\mathcal{W}} \mu_j = \prod_{j\in\mathcal{R}}(M+\mu_j)$, we further obtain

$$\begin{aligned}\Big|\mathrm{Cov}(\prod_{t\in\mathcal{S}} Y_t, \prod_{t\in\mathcal{T}} Y_t)\Big| \leq & K^2 \prod_{t\in\mathcal{R}} \gamma_t \prod_{j\in\mathcal{R}} (M+\mu_j) \Big\{ (u+v)! \Big\}^b \Psi(u,v)\rho(t_1 - s_u) \\ \leq & K^2 \Big\{ \gamma(M+\mu) \Big\}^{u+v} \Big\{ (u+v)! \Big\}^b \Psi(u,v)\rho(t_1 - s_u).\end{aligned}$$

Thus, (C.1) and (C.2) are satisfied by $Y_1, \ldots, Y_T$ with M replaced by $\gamma(M+\mu)$. This completes the proof. □

Using Lemma 11, we can remove the zero-mean requirement for $X_1, \ldots, X_T$ in Lemma 10. The next theorem summarizes Lemmas 10 and 11.

Theorem 25. *Let $X_1, \ldots, X_T$ be a sequence of random variables satisfying* (C.1)-(C.3). *Suppose $\mathbb{E}X_t = \mu_t$, and $|\mu_t| \leq \mu$ for $t = 1, \ldots, T$, where $\mu > 0$ is a constant. Let $S_T := \sum_{t=1}^T (X_t - \mu_t)$. Then, for any $t > 0$, we have*

$$\mathbb{P}(S_T \geq t) \leq \exp\Big\{ -\frac{t^2}{D_1 n + D_2 t^{(2a+2b+3)/(a+b+2)}} \Big\}. \tag{C.10}$$

Here D_1 and D_2 are constants defined by

$$D_1 = 2^{a+b+3}K^2(M+\mu)^2 L_1(K^2 \vee 2) \text{ and } D_2 = 2\Big\{(M+\mu)L_2(K^2 \vee 2)\Big\}^{1/(a+b+2)},$$

where a, b, K, M, L_1, L_2 are constants defined in (C.1)-(C.3)*.*

C.2 Supporting Lemma

Lemmas 12 - 14 are used in the proofs of Theorems 22 - 23. Lemmas 12 and 13 provide tail probabilities for related quantile-based statistics. Lemma 14 builds the connection between the tail probabilities of $\|\tilde{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}}\|_{\max}$ and $\|\hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}}\|_{\max}$.

Lemma 12. *Let $X \in \mathbb{R}$ be a random variable with distribution function F, and $X_1, \ldots, X_T$ be T realizations of X such that for any $\mathcal{S}, \mathcal{T} \subseteq \{1, \ldots, T\}$ with $\max(\mathcal{S}) \leq \min(\mathcal{T})$, we have*

$$\begin{aligned} &\Big|\mathbb{P}\Big(X_t \leq b, \forall t \in \mathcal{S} \cup \mathcal{T}\Big) - \mathbb{P}\Big(X_j \leq b, \forall j \in \mathcal{S}\Big)\mathbb{P}\Big(X_k \leq b, \forall k \in \mathcal{T}\Big)\Big| \qquad \text{(C.11)} \\ \leq &K^2\Psi(|\mathcal{S}|, |\mathcal{T}|)\rho\Big\{d(\mathcal{S}, \mathcal{T})\Big\}, \end{aligned}$$

where the sequence $\{\rho(t)\}_{t\geq 0}$ is nonincreasing and satisfies

$$\sum_{n=0}^{\infty}(n+1)^k\rho(n) \leq L_1 L_2^k (k!)^a, \ \forall\, k \geq 0, \qquad \text{(C.12)}$$

for some constants $K, L_1, L_2 > 0$ and $a \geq 0$. Then, for any $t > 0$ and $q \in (0,1)$, we have

$$\begin{aligned}&\mathbb{P}(|\hat{Q}(\{X_t\}; q) - Q(X; q)| \geq t) \\ &\leq \exp\Big(-\varphi\Big[F\Big\{F^{-1}(q) + t\Big\} - q - \frac{1}{T}\Big]\Big) + \exp\Big(-\varphi\Big[q - F\Big\{F^{-1}(q) - t\Big\}\Big]\Big),\end{aligned}$$

whenever we have $F\{F^{-1}(q) + t\} > q + 1/T$. Here the function φ is defined as

$$\varphi(x) := \frac{Tx^2}{D_1 + D_2 T^{(a+1)/(a+2)} x^{(2a+3)/(a+2)}}, \text{ for } x > 0, \tag{C.13}$$

where D_1 and D_2 are constants given by

$$D_1 = 2^{a+5} K^2 L_1 (K^2 \vee 2), \tag{C.14}$$

$$D_2 = 2\Big\{2L_2(K^2 \vee 2)\Big\}^{1/(a+2)}. \tag{C.15}$$

Proof. Let F_T be the empirical distribution function of $X_1, \ldots, X_T$ and $F_T^{-1}(q) = \hat{Q}(\{X_t\}; q)$. By the definition of $\hat{Q}(\cdot;\cdot)$ in (4.4), we have, for any $\epsilon \in [0,1]$,

$$\epsilon \leq F_T\{F_T^{-1}(\epsilon)\} \leq \epsilon + \frac{1}{T}. \tag{C.16}$$

By definition, we have

$$\mathbb{P}\Big\{\hat{Q}(\{X_t\}; q) - Q(X; q) \geq t\Big\} = \mathbb{P}\Big\{F_T^{-1}(q) - F^{-1}(q) \geq t\Big\}$$

$$\leq \mathbb{P}\Big[F_T\{F_T^{-1}(q)\} \geq F_T\{F^{-1}(q)+t\}\Big],$$

where the last inequality is because F_T is non-decreasing. By (C.16), we have

$$\mathbb{P}\Big\{\hat{Q}(\{X_t\};q) - Q(X;q) \geq t\Big\} \leq \mathbb{P}\Big[q + \frac{1}{T} \geq F_T\{t + F^{-1}(q)\}\Big].$$

By the definition of F_T, we further have

$$\begin{aligned}\mathbb{P}\Big\{\hat{Q}(\{X_t\};q) - Q(X;q) \geq t\Big\} &\leq \mathbb{P}\Big[\sum_{t=1}^{T} I\{X_t \leq F^{-1}(q)+t\} \leq nq+1\Big] \\ =\mathbb{P}\Big(\sum_{t=1}^{T}\Big[-I\{X_t \leq F^{-1}(q)+t\} &+ F\{F^{-1}(q)+t\}\Big] \geq T\Big[F\{F^{-1}(q)+t\} - q - \frac{1}{T}\Big]\Big).\end{aligned}$$

Using (C.11), we have

$$\mathrm{Cov}\Big[\prod_{t\in\mathcal{S}} I\Big\{X_t \leq F^{-1}(q)+t\Big\}, \prod_{t\in\mathcal{T}} I\Big\{X_t \leq F^{-1}(q)+t\Big\}\Big] \leq K^2\Psi(|\mathcal{S}|,|\mathcal{T}|)\rho\Big\{d(\mathcal{S},\mathcal{T})\Big\},$$

for any $\mathcal{S},\mathcal{T} \subseteq \{1,\ldots,T\}$ with $\max(\mathcal{S}) \leq \min(\mathcal{T})$. Thus, by Theorem 25, we have

$$\mathbb{P}\Big\{\hat{Q}(\{X_t\};q) - Q(X;q) \geq t\Big\} \leq \exp\Big(-\varphi\Big[F\{F^{-1}(q)+t\} - q - \frac{1}{T}\Big]\Big) \tag{C.17}$$

with function φ specified in (C.13). On the other hand, we have

$$\mathbb{P}\Big\{\hat{Q}(\{X_t\};q) - Q(X;q) \leq -t\Big\} = \mathbb{P}\Big\{F_T^{-1}(q) - F^{-1}(q) \leq -t\Big\}$$

$$\leq \mathbb{P}\Big[F_T\{F_T^{-1}(q)\} \leq F_T\{F^{-1}(q) - t\}\Big].$$

Using (C.16) again, we have

$$\begin{aligned}
&\mathbb{P}\Big\{\hat{Q}(\{X_t\}; q) - Q(X; q) \leq -t\Big\} \leq \mathbb{P}\Big[q \leq F_T\{F^{-1}(q) - t\}\Big] \\
=&\mathbb{P}\Big(\sum_{t=1}^{T}\Big[I\{X_t \leq F^{-1}(q) - t\} - F\{F^{-1}(q) - t\}\Big] \geq T\Big[q - F\{F^{-1}(q) - t\}\Big]\Big).
\end{aligned}$$

Thus, by Theorem 25, we have

$$\mathbb{P}\Big\{\hat{Q}(\{X_t\}; q) - Q(X; q) \leq -t\Big\} \leq \exp\Big(-\varphi\Big[q - F\{F^{-1}(q) - t\}\Big]\Big), \tag{C.18}$$

where the function φ is defined in (C.13). Combining (C.17) and (C.18) completes the proof. □

Lemma 13. *Let $X \in \mathbb{R}$ be a random variable. Denote by F and $\bar{F}$ the distribution functions of X and $|X - Q(X, 1/2)|$. Let $X_1, \ldots, X_T$ be T realizations of X satisfying* (C.11) *and* (C.12) *in Lemma 12. Then, for any $t > 0$, we have*

$$\begin{aligned}
&\mathbb{P}\Big(|\hat{\sigma}^{\mathrm{M}}(\{X_t\}_{t=1}^{T}) - \sigma^{\mathrm{M}}(X)| > t\Big) \\
\leq& 2\exp\Big(-\varphi\Big[F\Big\{F^{-1}(q) + \frac{t}{2}\Big\} - q - \frac{1}{T}\Big]\Big) + 2\exp\Big(-\varphi\Big[q - F\Big\{F^{-1}(q) - \frac{t}{2}\Big\}\Big]\Big) + \\
&\exp\Big(-\varphi\Big[\bar{F}\Big\{\bar{F}^{-1}(q) + \frac{t}{2}\Big\} - q - \frac{1}{T}\Big]\Big) + \exp\Big(-\varphi\Big[q - \bar{F}\Big\{\bar{F}^{-1}(q) - \frac{t}{2}\Big\}\Big]\Big).
\end{aligned}$$

whenever $F\{F^{-1}(q)+t/2\}-q>1/T$ *and* $\bar{F}\{\bar{F}^{-1}(q)+t/2\}-q>1/T$. *Here* φ *is defined in* (C.13).

Proof. We denote $\hat{m}:=\hat{Q}(\{X_t\}_{t=1}^T;1/2)$ and $m:=Q(X;1/2)$ to be the sample and population medians. By the definition of $\hat{\sigma}^{\mathrm{M}}(\cdot)$, we have

$$\begin{aligned}
&\mathbb{P}\Big\{\hat{\sigma}^{\mathrm{M}}\Big(\{X_t\}_{t=1}^T\Big)-\sigma^{\mathrm{M}}(X)>t\Big\}\\
=&\mathbb{P}\Big\{\hat{Q}\Big(\Big\{|X_t-\hat{m}|\Big\}_{t=1}^T;q\Big)-Q\Big(|X-m|;q\Big)>t\Big\}\\
\le&\mathbb{P}\Big\{\hat{Q}\Big(\Big\{|X_t-m|\Big\}_{t=1}^T;q\Big)+|\hat{m}-m|-Q\Big(|X-m|;q\Big)>t\Big\}\\
\le&\mathbb{P}\Big\{\hat{Q}\Big(\Big\{|X_t-m|\Big\}_{t=1}^T;q\Big)-Q\Big(|X-m|;q\Big)>\frac{t}{2}\Big\}+\mathbb{P}\Big(|\hat{m}-m|>\frac{t}{2}\Big).
\end{aligned}\tag{C.19}$$

On the other hand, using the same technique, we have

$$\begin{aligned}
&\mathbb{P}\Big\{\hat{\sigma}^{\mathrm{M}}\Big(\{X_t\}_{t=1}^T\Big)-\sigma^{\mathrm{M}}(X)<-t\Big\}\\
=&\mathbb{P}\Big\{\hat{Q}\Big(\Big\{|X_t-\hat{m}|\Big\}_{t=1}^T;q\Big)-Q\Big(|X-m|;q\Big)<-t\Big\}\\
\le&\mathbb{P}\Big\{\hat{Q}\Big(\Big\{|X_t-m|\Big\}_{t=1}^T;q\Big)-|\hat{m}-m|-Q\Big(|X-m|;q\Big)<-t\Big\}\\
\le&\mathbb{P}\Big\{\hat{Q}\Big(\Big\{|X_t-m|\Big\}_{t=1}^T;q\Big)-Q\Big(|X-m|;q\Big)<-\frac{t}{2}\Big\}+\mathbb{P}\Big(|\hat{m}-m|>\frac{t}{2}\Big).
\end{aligned}\tag{C.20}$$

Combining (C.19) and (C.20), we have

$$\begin{aligned}
&\mathbb{P}\Big\{|\hat{\sigma}^{\mathrm{M}}\Big(\{X_t\}_{t=1}^T\Big)-\sigma^{\mathrm{M}}(X)|>t\Big\}\\
\le&\mathbb{P}\Big\{\Big|\hat{Q}\Big(\Big\{|X_t-m|\Big\}_{t=1}^T;q\Big)-Q\Big(|X-m|;q\Big)\Big|>\frac{t}{2}\Big\}+2\mathbb{P}\Big(|\hat{m}-m|>\frac{t}{2}\Big).
\end{aligned}\tag{C.21}$$

Using Lemma 12, we have

$$\begin{aligned}
&\mathbb{P}\Big\{\Big|\hat{Q}\Big(\Big\{|X_t - m|\Big\}_{t=1}^{T}; q\Big) - Q\Big(|X - m|; q\Big)\Big| > \frac{t}{2}\Big\} \\
&\leq \exp\Big(-\varphi\Big[\bar{F}\Big\{\bar{F}^{-1}(q) + \frac{t}{2}\Big\} - q - \frac{1}{T}\Big]\Big) + \exp\Big(-\varphi\Big[q - \bar{F}\Big\{\bar{F}^{-1}(q) - \frac{t}{2}\Big\}\Big]\Big), \quad \text{(C.22)} \\
&\mathbb{P}\Big(|\hat{m} - m| > \frac{t}{2}\Big) \\
&\leq \exp\Big(-\varphi\Big[F\Big\{F^{-1}(q) + \frac{t}{2}\Big\}\Big] - q - \frac{1}{T}\Big) + \exp\Big(-\varphi\Big[q - F\Big\{F^{-1}(q) - \frac{t}{2}\Big\}\Big]\Big), \quad \text{(C.23)}
\end{aligned}$$

whenever $F\{F^{-1}(q) + t/2\} - q > 1/T$ and $\bar{F}\{\bar{F}^{-1}(q) + t/2\} - q > 1/T$. Combining (C.21), (C.22), and (C.23) leads to the desired result. □

Lemma 14. *For any $t \geq 0$, the solution $\tilde{\mathbf{R}}^{\mathrm{MAD}}$ to the optimization problem* (4.7) *satisfies*

$$\mathbb{P}\Big(\Big\|\tilde{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}}\Big\|_{\max} \geq t\Big) \leq \mathbb{P}\Big(\Big\|\hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}}\Big\|_{\max} \geq \frac{t}{2}\Big),$$

provided that $\mathbf{R}^{\mathrm{MAD}} \in \S_\lambda$.

Proof. When $\mathbf{R}^{\mathrm{MAD}} \in \S_\lambda$, it's feasible to optimization problem (4.7). This implies that

$$\Big\|\hat{\mathbf{R}}^{\mathrm{MAD}} - \tilde{\mathbf{R}}^{\mathrm{MAD}}\Big\|_{\max} \leq \Big\|\hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}}\Big\|_{\max}. \quad \text{(C.24)}$$

Thus, for any $t > 0$, we have

$$\mathbb{P}\Big(\Big\|\tilde{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}}\Big\|_{\max} \geq t\Big)$$

$$\leq \mathbb{P}\Big(\Big\| \tilde{\mathbf{R}}^{\mathrm{MAD}} - \hat{\mathbf{R}}^{\mathrm{MAD}} \Big\|_{\max} + \Big\| \hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}} \Big\|_{\max} \geq t \Big)$$
$$\leq \mathbb{P}\Big(\Big\| \hat{\mathbf{R}}^{\mathrm{MAD}} - \mathbf{R}^{\mathrm{MAD}} \Big\|_{\max} \geq \frac{t}{2} \Big).$$

Here the last inequality is due to (C.24). This completes the proof. □

Bibliography

Andersen, T. G. (2009). *Handbook of Financial Time Series*. Springer.

Andrews, D. W. (1984). Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21(4):930–934.

Antunes, A. M. C. and Rao, T. S. (2006). On hypotheses testing for the selection of spatio-temporal models. *Journal of Time Series Analysis*, 27(5):767–791.

Bai, J., Li, K., et al. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465.

Bai, J. and Liao, Y. (2012). Efficient estimation of approximate factor models via regularized maximum likelihood. *arXiv preprint arXiv:1209.5911*.

Bai, J. and Liao, Y. (2013). Statistical inferences using large estimated covariances for panel data and factor models. *arXiv preprint arXiv:1307.2662*.

Bai, J. and Shi, S. (2011). Estimating high dimensional covariance matrices and its applications. *Annals of Economics and Finance*, 12(2):199–215.

BIBLIOGRAPHY

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Bartzokis, G., Beckson, M., Lu, P. H., Nuechterlein, K. H., Edwards, N., and Mintz, J. (2001). Age-related changes in frontal and temporal lobe volumes in men: a magnetic resonance imaging study. *Archives of General Psychiatry*, 58(5):461–465.

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Best, M. J. and Grauer, R. R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *Review of Financial Studies*, 4(2):315–342.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739.

BIBLIOGRAPHY

Blakemore, S.-J. (2012). Imaging brain development: the adolescent brain. *Neuroimage*, 61(2):397–406.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2(2):107–144.

Braun, U., Plichta, M. M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., et al. (2012). Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. *Neuroimage*, 59(2):1404–1412.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer.

Cai, T., Liu, W., and Luo, X. (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Chanda, K. C. (1974). Strong mixing properties of linear stochastic processes. *Journal of Applied Probability*, 11(2):401–408.

Chen, G., Glen, D. R., Saad, Z. S., Paul Hamilton, J., Thomason, M. E., Gotlib, I. H., and Cox, R. W. (2011a). Vector autoregression, structural equation modeling, and their syn-

thesis in neuroimaging data analysis. *Computers in Biology and Medicine*, 41(12):1142–1155.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021.

Chen, Y., Wiesel, A., and Hero, A. O. (2011b). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107.

Chopra, V. K. and Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management*, 19(2):6–11.

Couillet, R. and McKay, M. R. (2014). Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, 131:99–120.

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397.

BIBLIOGRAPHY

Dedecker, J. and Prieur, C. (2004). Coupling for τ-dependent sequences and applications. *Journal of Theoretical Probability*, 17(4):861–885.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84(2):313–342.

Doukhan, P. and Neumann, M. H. (2007). Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, 117(7):878–903.

Doukhan, P. and Neumann, M. H. (2008). The notion of ψ-weak dependence and its applications to bootstrapping time series. *Probability Surveys*, 5:146–168.

Dowd, K. (2007). *Measuring Market Risk*. Wiley.

Drton, M. and Perlman, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 22(3):430–449.

Drton, M. and Perlman, M. D. (2008). A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200.

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., Barber, A. D., Joel, S., Pekar, J. J., Mostofsky, S. H., et al. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6(61):1–9.

BIBLIOGRAPHY

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356.

Fan, J., Liao, Y., and Mincheva, M. (2013a). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Fan, J., Liao, Y., and Shi, X. (2013b). Risks of large portfolios. *arXiv preprint arXiv:1302.0926*.

Fan, J., Zhang, J., and Yu, K. (2012a). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

Fan, J., Zhang, J., and Yu, K. (2012b). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

Fang, K.-T., Kotz, S., and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friston, K. (2011). Functional and effective connectivity: a review. *Brain Connectivity*, 1(1):13–36.

BIBLIOGRAPHY

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., Paus, T., Evans, A. C., and Rapoport, J. L. (1999). Brain development during childhood and adolescence: a longitudinal MRI study. *Nature Neuroscience*, 2(10):861–863.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124.

Gorodetskii, V. (1978). On the strong mixing property for linear sequences. *Theory of Probability and Its Applications*, 22(2):411–413.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press, Oxford.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Han, F. and Liu, H. (2013a). Principal component analysis on non-Gaussian dependent

data. In *Proceedings of the 30th International Conference on Machine Learning*, pages 240–248.

Han, F. and Liu, H. (2013b). Transition matrix estimation in high dimensional time series. In *Proceedings of the 30th International Conference on Machine Learning*, pages 172–180.

Han, F., Liu, H., and Caffo, B. (2013). Sparse median graphs estimation in a high dimensional semiparametric model. *arXiv preprint arXiv:1310.3223*.

Han, F., Lu, J., and Liu, H. (2014). Robust scatter matrix estimation for high dimensional distributions with heavy tails. Technical report, Johns Hopkins University.

Harrison, L., Penny, W. D., and Friston, K. (2003). Multivariate autoregressive modeling of fMRI time series. *Neuroimage*, 19(4):1477–1491.

Høst, G., Omre, H., and Switzer, P. (1995). Spatial interpolation errors for monitoring data. *Journal of the American Statistical Association*, 90(431):853–861.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1683.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall.

Jones, R. H. and Zhang, Y. (1997). Models for continuous stationary space-time processes. In *Modelling Longitudinal and Spatially Correlated Data*, pages 289–298. Springer.

BIBLIOGRAPHY

Kallabis, R. S. and Neumann, M. H. (2006). An exponential inequality under weak dependence. *Bernoulli*, 12(2):333–350.

Kallberg, J. G. and Ziemba, W. T. (1984). Mis-specifications in portfolio selection problems. In *Risk and Capital*, pages 74–87. Springer.

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010). Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123.

Kolar, M. and Xing, E. P. (2009). Sparsistent estimation of time-varying discrete Markov random fields. *arXiv preprint arXiv:0907.2337*.

Kontorovich, L. A., Ramanan, K., et al. (2008). Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Lauritzen, S. L. (1996). *Graphical Models*, volume 17. Oxford University Press.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Ledoit, O. and Wolf, M. (2004a). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

BIBLIOGRAPHY

Ledoit, O. and Wolf, M. (2004b). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012a). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Liu, H., Han, F., and Zhang, C.-H. (2012b). Transelliptical graphical models. In *Advances in Neural Information Processing Systems*, pages 809–817.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*, pages 1432–1440.

Liu, W. and Luo, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. *arXiv preprint arXiv:1203.3896*.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Majda, A. and Wang, X. (2006). *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*. Cambridge University Press.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

BIBLIOGRAPHY

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pages 273–292. Institute of Mathematical Statistics.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361.

Mohri, M. and Rostamizadeh, A. (2010). Stability bounds for stationary ϕ-mixing and β-mixing processes. *The Journal of Machine Learning Research*, 11:789–814.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097.

BIBLIOGRAPHY

Nobre, A. A., Sansó, B., and Schmidt, A. M. (2011). Spatially varying autoregressive processes. *Technometrics*, 53(3):310–321.

Pan, J. and Yao, Q. (2008). Modeling multiple time series via common factors. *Biometrika*, 95(2):365–379.

Pang, H., Liu, H., and Vanderbei, R. (2013). The fastclime package for linear programming and constrained L1-minimization approach to sparse precision matrix estimation in R. *CRAN*.

Penny, W., Ghahramani, Z., and Friston, K. (2005). Bilinear dynamical systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):983–993.

Pham, T. D. and Tran, L. T. (1985). Some mixing properties of time series models. *Stochastic Processes and their Applications*, 19(2):297–303.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., et al. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Purdon, P. L. and Weisskoff, R. M. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human Brain Mapping*, 6(4):239–249.

Qiu, H., Han, F., Liu, H., and Caffo, B. (2015). Joint estimation of multiple graphical

models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (in press)*.

Rachev, S. T. (2003). *Handbook of Heavy Tailed Distributions in Finance*. Elsevier.

Rachev, S. T., Menn, C., and Fabozzi, F. J. (2005). *Fat-tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing*. Wiley.

Rao, S. S. (2008). Statistical analysis of a spatio-temporal model with location-dependent parameters and a test for spatial stationarity. *Journal of Time Series Analysis*, 29(4):673–694.

Roger, H. and Charles, R. J. (1994). *Topics in Matrix Analysis*. Cambridge University Press.

Rogers, B. P., Katwal, S. B., Morgan, V. L., Asplund, C. L., and Gore, J. C. (2010). Functional MRI and multivariate autoregressive models. *Magnetic Resonance Imaging*, 28(8):1058–1065.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

BIBLIOGRAPHY

Sancetta, A. (2008). Sample covariance shrinkage for high dimensional dependent data. *Journal of Multivariate Analysis*, 99(5):949–967.

Schmidt, R. (2002). Tail dependence for elliptically contoured distributions. *Mathematical Methods of Operations Research*, 55(2):301–327.

Shaw, P., Kabani, N. J., Lerch, J. P., Eckstrand, K., Lenroot, R., Gogtay, N., Greenstein, D., Clasen, L., Evans, A., Rapoport, J. L., et al. (2008). Neurodevelopmental trajectories of the human cerebral cortex. *The Journal of Neuroscience*, 28(14):3586–3594.

Smith, G. D. (1978). *Numerical Solution of Partial Differential Equations: Finite Difference Methods*. Clarendon Press: Oxford.

Sølna, K. and Switzer, P. (1996). Time trend estimation for a geographic region. *Journal of the American Statistical Association*, 91(434):577–589.

Song, L., Kolar, M., and Xing, E. P. (2009a). KELLER: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128–i136.

Song, L., Kolar, M., and Xing, E. P. (2009b). Time-varying dynamic bayesian networks. In *Advances in Neural Information Processing Systems*, pages 1732–1740.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

BIBLIOGRAPHY

Storvik, G., Frigessi, A., and Hirst, D. (2002). Stationary space-time Gaussian fields and their time autoregressive representation. *Statistical Modelling*, 2(2):139–161.

Tasaki, H. (2009). Convergence rates of approximate sums of Riemann integrals. *Journal of Approximation Theory*, 161(2):477–490.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., and Canales-Rodríguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):969–981.

Van De Geer, S. and Van De Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press, Cambridge.

Van Dijk, K. R., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage*, 59(1):431–438.

Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222.

Wang, Z., Han, F., and Liu, H. (2013). Sparse principal component analysis for high dimensional multivariate time series. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 48–56.

BIBLIOGRAPHY

Wold, H. (1938). A study in the analysis of stationary time series. *Journal of the Royal Statistical Society*, 102(2):295–298.

Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *Neuroimage*, 14(6):1370–1386.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154.

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *The Annals of Statistics*, 40(1):466–493.

Xu, M. and Shao, H. (2012a). Solving the matrix nearness problem in the maximum norm by applying a projection and contraction method. *Advances in Operations Research*, 2012:1–15.

Xu, M. H. and Shao, H. (2012b). Solving the matrix nearness problem in the maximum norm by applying a projection and contraction method. *Advances in Operations Research*, 2012:1–15.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286.

BIBLIOGRAPHY

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *The Journal of Machine Learning Research*, 13:1059–1062.

Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319.

BIBLIOGRAPHY

CURRICULUM VITAE

HUITONG QIU

hqiu7@jhu.edu

615 N. Wolfe St. E3032

Baltimore, MD 21205

Date of Birth: December 26th, 1988

Place of Birth: Hebei, China

EDUCATION

2011 - 2016 **Johns Hopkins Bloomberg School of Public Health**, Baltimore, MD

Ph.D. in Biostatistics

Thesis title: *Statistical Methods and Theory for Analyzing High Dimensional Time Series*

Advisor: Prof. Brian Caffo

2007 - 2011 **Fudan University**, Shanghai, China

B.S. in Mathematics

PROFESSIONAL EXPERIENCE

06/2015 - 08/2015 **Summer Associate**

CURRICULUM VITAE

	Goldman Sachs, New York City, NY
06/2014 - 08/2014	**Summer Intern**
	AT&T Labs Research, Middletown, NJ

HONORS AND AWARDS

2016	Joseph Zeger Conference Travel Award
2015	Neural Information Processing Systems (NIPS) Student Travel Award
2014	Student/Young Researcher Paper Award, Risk Analysis Section, American Statistical Association (ASA)
2014	Eastern North American Region (ENAR) Distinguished Student Paper Award
2010	Liao Kaiyuan Scholarship (top 3%)
2009	National 1st Prize of China Undergraduate Mathematical Contest in Modeling

---

PUBLICATIONS

PUBLISHED/SUBMITTED

Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. Robust Portfolio Optimization. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

CURRICULUM VITAE

Huitong Qiu, Fang Han. Robust Estimation of High Dimensional Heavy-tailed Vector Autoregressive Processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1843–1851, 2015.

Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. Joint Estimation of Multiple Graphical Models from High Dimensional Time Series. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 87(2), pp. 478 – 504, 2015.

Michael Rosenblum, Tianchen Qian, Yu Du, and **Huitong Qiu**. Adaptive Enrichment Designs for Randomized Trials with Delayed Endpoints, using Locally Efficient Estimators to Improve Precision. Accepted by *Biostatistics*, 2015.

WORKING PAPERS

Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. A Theory of Kolmogorov Dependence with Applications to Scatter Matrix Estimation. 2015.

Fang Han, **Huitong Qiu**, Brian Caffo. On the Impact of Dimension Reduction on Graphical Structures. 2015.

Shaojie Chen, Lei Huang, **Huitong Qiu**, Mary Beth Nebel, Stewart Mostofsky, James Pekar, Martin Lindquist, Ani Eloyan, Brian Caffo. A Parallel Group Independent Component Analysis Algorithm for Massive fMRI Data Analysis. 2015.

PRESENTATIONS

CURRICULUM VITAE

2015 Robust Portfolio Optimization. The 29th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada

2014 Robust Portfolio Optimization under High Dimensional Heavy-tailed Time Series. Joint Statistical Meeting (JSM), Boston, MA, USA

2014 Joint Estimation of Multiple Graphical Models from High Dimensional Time Series. Eastern North American Region (ENAR) Spring Meeting, Baltimore, MD, USA

TEACHING

2016 Statistical Methods in Public Health III, Graduate, 140.623, Profs. Marie Diener West and John McGready

2015 Statistical Reasoning in Public Health I-II, Graduate, 140.611-612, Prof. John McGready

2015 Statistical Methods in Public Health III-IV, Graduate, 140.623-624, Profs. Marie Diener West and John McGready

2014 Statistical Reasoning in Public Health I-II, Graduate, 140.611-612, Prof. John McGready

2014 Introduction to Statistical Theory I-II, Graduate, 140.673-674, Prof. Constantine Frangakis

CURRICULUM VITAE

2013 Statistical Reasoning in Public Health I-II, Graduate, 140.611-612, Prof. John McGready

2013 Statistical Methods in Public Health III-IV, Graduate, 140.623-624, Porfs. Marie Diener West and Karen Bandeen-Roche

2012 Statistical Methods in Public Health I-II, Graduate, 140.621-622, Porfs. Marie Diener West and Karen Bandeen-Roche